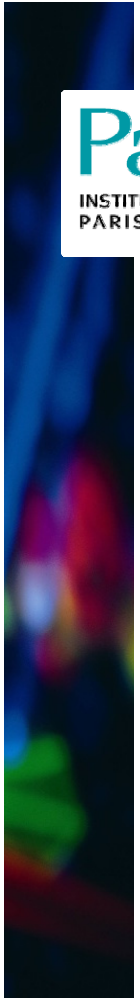
Méthodologie d'extraction de connaissances spatio-temporelles par fouille de données pour l'analyse de comportements à risques : application à la surveillance maritime

Bilal Idiri



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

MINES ParisTech

École doctorale n° 432 : Sciences des Métiers de l'Ingénieur

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

l'École nationale supérieure des mines de Paris

Spécialité " Sciences et Génie des Activités à Risques "

présentée et soutenue publiquement par

Bilal IDIRI

le 17 décembre 2013

Méthodologie d'extraction de connaissances spatio-temporelles par fouille de données pour l'analyse de comportements à risques - Application à la surveillance maritime

Directeur de thèse : **Aldo NAPOLI**

Jury

M. Thomas DEVOGELE, Professeur, Université François Rabelais de Tours/Laboratoire d'informatique	Rapporteur
M. Alain BOUJU, Maître de Conférences HDR, Université de La Rochelle/Laboratoire l3i	Rapporteur
Mme Karine ZEITOUNI, Professeur, Université de Versailles Saint-Quentin-en-Yvelines/Laboratoire PRISM	Examinateur
M. Cyril RAY, Maître de Conférences, l'Ecole Navale/IRENav	Examinateur
M. Franck GUARNIERI, Directeur de recherche HDR, MINES ParisTech/CRC	Examinateur
M. Aldo NAPOLI, Chargé de recherche HDR, MINES ParisTech/CRC	Examinateur

T H È S E

MINES ParisTech
Centre de recherche sur les Risques et les Crises
CS 10207
06904 Sophia Antipolis Cedex

A ce que j'ai de plus cher, Amel, ma petite Aya, mon beau-père Belkacem et mes parents, Amar et Malika.

Merci pour votre amour et vos encouragements.

Remerciements

Je tiens à remercier les membres du jury de m'avoir fait l'honneur de leur participation. Je remercie Thomas Devogele (Université François Rabelais) et Alain Bouju (Université de La Rochelle) d'avoir accepté d'être rapporteurs de cette thèse. Karine Zeitouni (Université de Versailles), Cyril Ray (Ecole Navale) et Franck Guarnieri (MINES ParisTech) d'avoir bien voulu être examinateurs de ce travail.

Un grand merci pour tous les membres du Centre de recherche sur les Risques et les Crises (CRC) et particulièrement à son Directeur, Franck Guarnieri, pour m'avoir permis de vivre cette belle et enrichissante expérience d'apprenti-chercheur au sein de son équipe ! à Gabriel Vatin et Xavier Chaze pour leur amitié. Merci à eux d'avoir relu, annoté et corrigé ce mémoire. Je passe une dédicace spéciale à tous mes collègues que j'ai eu le plaisir de côtoyer durant ces trois années à Sophia Antipolis. Ces années ont été riches d'enseignement grâce à nos différents échanges. Merci à Sandrine Renaux, Stéphanie Garnier, Myriam Perrault Lavigne, Sylvie Michel pour leur disponibilité et à Eric Rigaud, Valérie Godfrin, Jean-Luc Wybo pour leurs précieux conseils.

J'en profite pour remercier aussi Alain Orengo, directeur de la société ITE et son équipe, de m'avoir apporté de l'aide dans la conception du prototype ShipMine, à leurs conseils et au suivi régulier de mon travail. Je tiens à adresser un remerciement spécial à Adnan El Moussawi qui a travaillé au cours de son stage sur le développement de ShipMine.

Mes derniers remerciements vont à mon directeur de thèse Aldo Napoli (MINES ParisTech) pour sa disponibilité et sa confiance. Je le remercie grandement pour l'autonomie qu'il m'a accordée durant ces quelques années passées au sein du CRC. J'ai aussi apprécié ses qualités humaines, merci pour tout Aldo.

Table des matières

Liste des figures

Liste des tables

Liste des algorithmes

Introduction

Contexte

L'activité maritime : un contexte générateur de risques

La mer représente 75% de la surface du globe terrestre, l'être humain l'utilise pour se déplacer, pêcher et transporter des marchandises lourdes à moindre coût. L'activité maritime qui y est engendrée par l'utilisation de la mer est l'un des facteurs essentiels du développement économique et du rapprochement des territoires éloignés. Elle évolue sur un espace à la fois spatio-temporel par son contexte physique mais aussi stratégique par son emploi. Selon un rapport de la conférence des Nations Unies sur le commerce et le développement (CNUCED 2009), l'économie maritime représente 90% des échanges internationaux avec 80% du transport d'énergie. On recense 1,19 milliard de tonnes de port en lourd[1] (TPL) en 2009 avec une croissance de 6,7% par rapport à 2008 malgré la récession économique. Ce qui est une preuve de l'importance de ce secteur. Au niveau de l'Union Européenne (UE), la flotte commerciale a connu une augmentation de 80% de sa capacité durant ces 30 dernières années. L'Agence Internationale de l'Energie (AIE) prévoit une augmentation de 50 % d'ici 2020 de la demande européenne de gaz ce qui pourrait envoyer encore à la hausse la capacité de la flotte commerciale.

L'activité maritime évolue dans un espace maritime complexe, caractérisé par sa dimension internationale due à l'existence d'espaces internationaux libres et de plusieurs états côtiers ayant chacun leur propre réglementation maritime. La plupart des états côtiers possèdent une multitude d'acteurs intervenant en mer comme les centres de surveillance (Centre Régional Opérationnel de Surveillance et de Sauvetage en France, etc.), d'enquête maritimes (*BEAmer*[2] en France, MAIB[3] en Angleterre, etc.), la Marine nationale, la gendarmerie, les douanes et les gardes côtes. Ces acteurs assurent des missions variées comme la lutte antipollution, la surveillance du trafic, le contrôle des pêches, la sécurité de la navigation maritime, la sûreté, la répression des trafics illicites et le sauvetage de personnes. Les missions sont conduites la plupart du temps dans des périmètres restreints, rattachés à des découpages administratifs. D'autres sont menées en

[1] Nombre de tonnes qu'un navire peut transporter.
[2] http://www.beamer-france.org/index.php
[3] http://www.maib.gov.uk/home/index.cfm

haute mer dans des espaces vastes comme celles effectuées par la Marine nationale qui a pour mission de traiter les menaces au plus loin des côtes et au plus près de leur source[4]. Les navires doivent êtres blanchis[5] au plutôt avant leur accostage pour une meilleure sûreté (Prati 2010).

Les découpages administratifs utilisés permettent de subdiviser l'espace maritime en zones virtuelles partant des côtes vers la haute mer. On trouve parmi ces découpages, la mer territoriale[6], les Zones Economiques Exclusives (ZEE), les Zones de pêche, les Zones de Protection de la Pêche (ZPP) et les Zones de Protection Ecologique (ZPE). L'objectif de ces découpages est de permettre une meilleure appropriation des espaces maritimes, faciliter leur gestion, leur surveillance et éviter leur utilisation abusive. Du fait de l'importance liée à la détermination de ces zones, leur nombre augmente rapidement. En Méditerranée, ce nombre s'est multiplié en une décennie avec différentes nominations, dimensions et caractéristiques (Heredia 2009). Ces zones permettent à un Etat côtier de définir des droits exclusifs d'exploitation, d'exploration et de gestion des ressources sur elles et aux autres Etats, le droit de naviguer, de survoler et de poser des réseaux sous-marins (pipelines, câbles, etc.) si cela ne cause pas de tort (Bellayer Roille 2011).

L'évolution de l'activité maritime dans cet espace ouvert, l'augmentation du trafic maritime mondial, sa densité dans certaines zones, les différentes réglementations entre les états côtiers et les enjeux économiques et politiques de cette activité forment un contexte générateur de risques sur les navires, les personnes, les états et l'environnement (écologique et infrastructures portuaires). Nous adoptons pour la suite de ce manuscrit, la définition de Boisson (Boisson 1998) qui définit le risque comme l'éventualité d'un événement pouvant provoquer des conséquences dommageables. Dans le domaine maritime, les risques peuvent être liés à la sécurité (échouement, naufrage, collision, etc.) ou à la sûreté maritime (attaque terroriste, immigration illégale, trafics de biens illicites, etc.). Ceux liés à la sécurité maritime sont causés par un dysfonctionnement ou une activité mettant en danger les navires, les personnes et l'environnement alors que les risques liés à la sûreté sont dus à un acte illicite à l'encontre du navire, des personnes et

[4] http://www.cluster-maritime.fr/article.php?id=20

[5] Vérifiés qu'ils ne présentent pas de menaces.

[6] C'est une zone sous souveraineté d'un Etat côtier bordant la côte d'une largeur ne dépassant pas les 12 milles s'il n'y a pas de médiation entre deux états voisins.

des infrastructures portuaires. Quelques chiffres sont présentés ci-après pour donner un aperçu des risques qui pèsent sur cette activité maritime.

Pour faire face aux risques, plusieurs stratégies peuvent être distinguées à savoir la prévention, l'acceptation, la réduction et le transfert du risque à des compagnies d'assurance ou en sous-traitant l'activité à risque (Cougnaud 2007). La stratégie de réduction des risques est la plus classique, elle commence par identifier les risques et leurs facteurs pour mettre en place des actions de prévention. La prévention consiste à poser des mesures pour limiter l'apparition de scénarios négatifs ou les éviter en suspendant par exemple l'activité à risque. Dans la suite de ce travail, l'attention est portée sur la réduction des risques maritimes et le recours à des moyens de surveillance pour assurer la sécurité et la sûreté maritimes.

Les moyens mis en œuvre pour assurer la sécurité et la sûreté

Les moyens mis en œuvre aujourd'hui pour assurer la sécurité et la sûreté maritime peuvent être scindés en trois groupes : les moyens réglementaires, organisationnels et technologiques.

1. Moyens réglementaires

Sur le plan réglementaire, des efforts ont été ressentis pour assurer la sécurité et la sûreté maritime. Les Nations Unis par exemple ont créé en 1948 un organisme appelé OMI (Organisation Maritime Internationale[7]) qui a comme principales missions aujourd'hui de rendre plus sûr, efficace, durable et écologique le secteur maritime. Pour assurer ses missions, l'OMI adopte des règlementations de sécurité comme l'instauration de la double coque pour les navires pétroliers après le naufrage de l'ERIKA[8], la reconnaissance officielle des cartes de navigation électroniques (ENC[9]) et l'autorisation des états côtiers à enlever les épaves des navires se trouvant dans leur ZEE avec une

[7] Organisme international consacré exclusivement à la mise en œuvre de mesures visant à renforcer la sécurité, la sûreté, l'efficacité de la navigation et la prévention de la pollution en mer (OMI 2009)(OMI 2009).

[8] Le naufrage de l'ERIKA en image : http://tempsreel.nouvelobs.com/galeries-photos/planete/20100330.OBS1469/en-images-le-naufrage-de-l-erika-en-decembre-1999.html Images

[9] Les ENC sont l'équivalant électronique des cartes marines papier qui répertorient les descriptions détaillées des objets marins dans une base de données pour permettre leur exploitation par des systèmes informatiques.

indemnisation des opérations par les propriétaires des navires ou de leur assurance[10]. Avant l'ERIKA, le naufrage du Titanic[11] en 1916 a incité à une première convention pour la sauvegarde de la vie humaine en mer (SOLAS). Cette convention dresse un ensemble de règles et d'exigences à respecter pour assurer la sécurité, la sûreté du navire et de son équipage. Aujourd'hui SOLAS en est à sa cinquième version.

2. Moyens organisationnels

Sur le plan organisationnel, des organismes en charge de la sécurité et de la sûreté en mer ont été créés. Ces organismes ont souvent besoin d'une autorité coordinatrice et d'une approche globale pour le partage d'informations en matière de surveillance maritime (Bodewig et al., 2009). Le réseau de surveillance Français représenté sur la Figure 0-1, compte aujourd'hui 59 sémaphores, 4 Centres Opérationnels de la Marine (CO Marine), 7 CROSS[12], 4 Centre Opérationnels des douanes (CO Douanes), une fonction de garde côtes (CoFGC), etc.

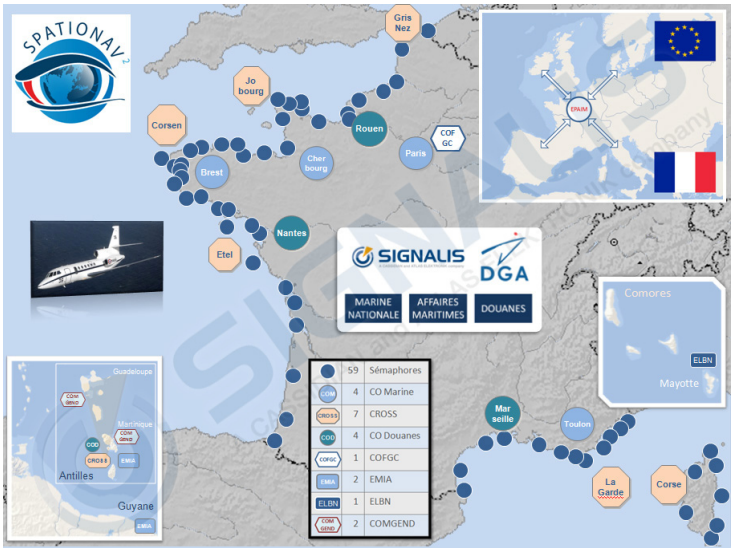


Figure 0-1 : Les moyens organisationnels maritimes mis en œuvre en France (http://www.signalis.com).

Dans le cadre de l'amélioration de leurs moyens organisationnels, différentes restructurations des organismes en charge de la sécurité et sûreté ont été effectuées.

[10] http://www.actu-environnement.com/ae/news/enlevement-epaves-ratification-convention-internationale-nairobi-19035.php4

[11] Le naufrage du Titanic en image : http://tempsreel.nouvelobs.com/galeries-photos/le-titanic-100-ans-apres/20120413.OBS6115/en-images-titanic-du-naufrage-a-la-legende.html

[12] Centres Régionaux Opérationnels de Surveillance et de Sauvetage : assurent la sécurité maritime dans le cadre de l'action de l'Etat en mer.

Prenons l'exemple de la création de la fonction de Garde-côtes dans l'objectif d'assurer des missions d'observation, d'analyse des flux maritimes, être le point d'entrée des coopérations internationales en termes de situations maritimes et information du gouvernement pour l'adaptation de leurs priorités[13]. Le Centre de Garde-côtes (CoFGC) est opérationnel depuis le 20 septembre 2010. Des unités de la Gendarmerie maritime appelées Pelotons de Sûreté Maritime et Portuaire (PSMP[14]) ont été aussi créées pour assurer la sûreté des navires, des approches maritimes et des installations portuaires contre les attaques terroristes.

Même si les outils règlementaires et organisationnels sont biens établis, leur application demande des moyens technologiques adaptés.

3. Moyens technologiques

Sur le plan technologique, des moyens de surveillance sont utilisés pour améliorer la sécurité et la sûreté maritime. Parmi les moyens de surveillance, nous pouvons distinguer les systèmes d'aide à la navigation et les systèmes de surveillance maritimes. Ces systèmes sont composés d'une infrastructure de détection (AIS, radar, etc.) permettant de capturer et transmettre les données de géolocalisation de navires et d'un système de traitement d'informations pour traiter, stocker et restituer les dernières informations sur les navires via des dispositifs d'affichage.

Avant de découvrir les systèmes d'aide à la navigation et les systèmes de surveillance, nous allons présenter les infrastructures de détection qui compose ces systèmes et les alimentent en données de déplacement de navires.

a. Infrastructures de détection

Les moyens de détection comme l'AIS et le radar sont très utilisés pour la surveillance maritime. Ces moyens permettent aux navires et aux systèmes de surveillance de détecter les positions des navires et leur déplacement.

L'AIS est composé d'une antenne de réception, de transmission VHF (Very High Frequency), d'un système de géolocalisation (GPS), un capteur de direction, un capteur

[13] http://www.sgmer.gouv.fr/Fonction-garde-cotes.html
[14] http://www.gouvernement.fr/gouvernement/pelotons-de-surete-maritime-et-portuaire-psmp

de vitesse, un écran de contrôle et d'autres composantes. Pour des raisons d'optimisation, la fréquence de synchronisation ou de transmission d'informations AIS est relative à la vitesse du navire, à son changement de direction et au type d'informations échangées. Parmi les informations échangées, on peut citer le code unique MMSI[15], la position, le cap, la vitesse, le tirant d'eau (partie immergée du navire), le port de rattachement, la cargaison, le temps de transmission et le temps d'arrivée estimé du navire. Les informations de type statique comme le port de rattachement et la cargaison ont un délai de synchronisation plus long que par exemple la vitesse et le cap qui sont de type dynamique.

Les radars sont fréquemment utilisés pour l'aide à la navigation. Ils sont composés d'un émetteur et récepteur électromagnétique permettant de détecter les objets voisins, de calculer leur position, leur vitesse par la réception d'ondes émises et réfléchies par les objets cibles (navires, obstacles, etc.). La position est calculée par rapport au temps de réflexion de l'onde et la position angulaire de l'émetteur. La vitesse est relative au décalage de fréquence de l'onde réfléchie.

Les AIS et les radars sont complémentaires. En effet, les navires non équipés de l'AIS sont détectés par radar dès qu'ils sont à la portée des ondes radars. Cependant si ces ondes n'arrivent pas à les détecter à cause de la portée de l'onde, de la présence d'obstacles empêchant leur propagation ou de la petite taille du navire, le transpondeur AIS permet de détecter les navires qui en sont équipés comme le montre la Figure 0-2.

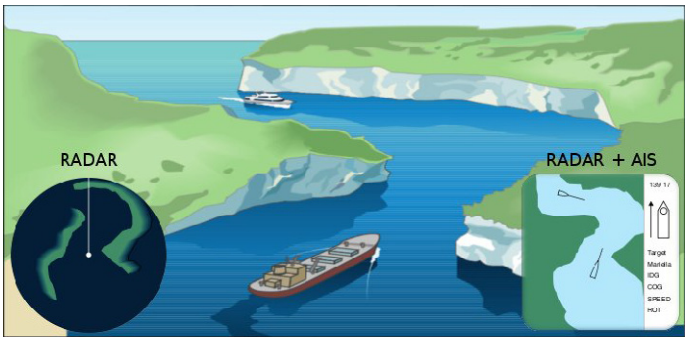


Figure 0-2 : La complémentarité entre le signal radar et l'AIS (Etienne 2011).

[15] Maritime Mobile Service Identity est un numéro unique d'équipements radio mis sur un navire. Il est utilisé pour identifier les navires exploitant ces équipements.

L'inconvénient de l'AIS et des radars conventionnels réside dans la portée du signal qui est entre 30 et 50 miles (la portée radio) des côtes : au-delà de cette limite, les navires ne sont donc plus détectés.

b. Les systèmes d'aide à la navigation

Les systèmes d'aide à la navigation sont des systèmes qui équipent les navires de radar et d'un logiciel conforme à la norme des systèmes de visualisation des cartes électroniques et d'information (ECDIS[16]) (Noyon 2007).

Ces systèmes sont couramment utilisés pour aider les marins en leur fournissant des informations précises sur l'environnement de navigation. Comme on le voit sur la Figure 0-3, la bathymétrie, les côtes, les obstacles, la réglementation, les navires à proximité et d'autres informations sont affichées sur ces systèmes. Parmi les systèmes les plus communément utilisés, citons NavTrack, Marine GIS et ex-Trem.

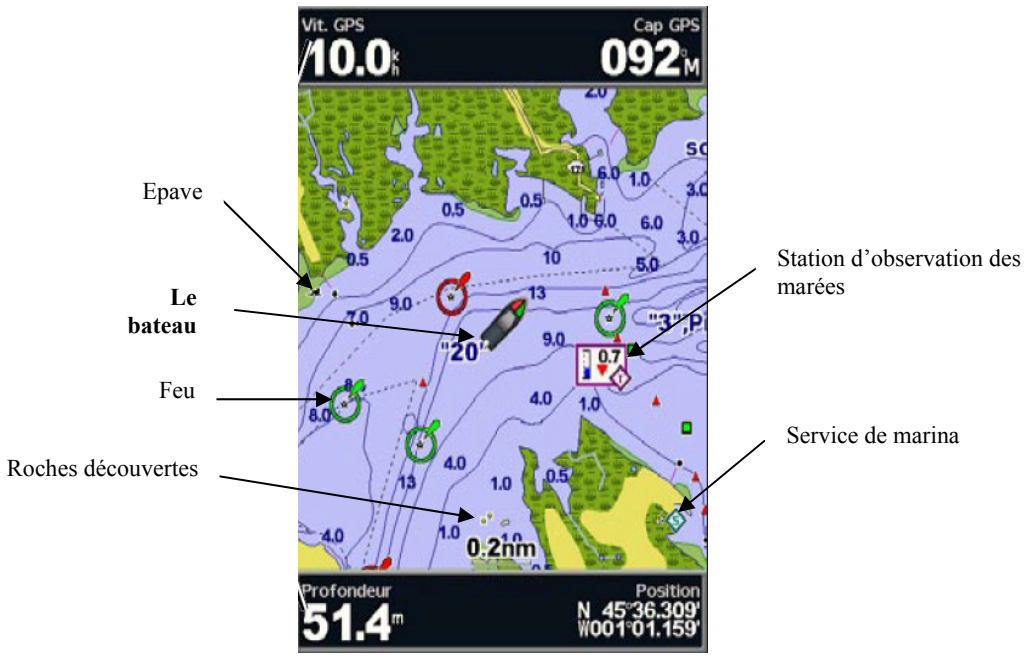


Figure 0-3 : Capture écran de l'interface d'un système d'aide à la navigation (source Garmin[17]).

[16] Electronic Chart Display and Information System, sont des systèmes de visualisation de positions de mobiles sur une cartographie électronique selon la norme de l'OMI.

[17] https://www.garmin.com/

c. Les systèmes de surveillance du trafic maritime

Le développement des systèmes de surveillance maritime comme SPATIONAV, ICSS, SIVE et SYTAR (Morel 2009) est le fruit des avancées technologiques en géolocalisation, en télécommunication, en systèmes embarqués et en cartes numériques. Ces systèmes sont des systèmes centralisés permettant aux centres de surveillance le suivi sur un dispositif d'affichage du trafic maritime en quasi temps-réel. Ces systèmes sont basés sur les informations AIS et pistes radar affichées sur des cartes ENC intégrées aux systèmes pour permettre l'analyse temps-réel des comportements de navires.

La majorité des pays de l'Union Européenne se sont équipés de systèmes de surveillance maritime après la directive 2002/59/CE du Parlement Européen et du Conseil du 27 juin 2002. La marine française a adopté un système de surveillance maritime appelé SPATIONAV qui permet le suivi du trafic maritime au large et dans tous les ports de l'hexagone métropolitain. La Direction Générale de l'Armement (DGA) en collaboration avec la Marine, conduit un projet d'amélioration de SPATIONAV pour bien répondre aux objectifs évolutifs du secteur de la surveillance maritime (Britz 2011). Comme SPATIONAV, la plupart des systèmes de surveillance maritime existant présentent des limites qu'il est possible de contrer en améliorant les infrastructures de détection et les systèmes de traitement d'informations.

L'évolution spatio-temporelle des navires sous-entend qu'il existe deux échelles d'observation du trafic maritime : une échelle spatiale qui peut être locale ou globale et une échelle temporelle. L'échelle locale se focalise sur la surveillance de l'espace autour du navire, d'une plateforme ou d'un port (Figure 0-4) et l'échelle globale réunit plusieurs espaces maritimes.

Selon cette définition, il est possible de dire que les systèmes d'aide à la navigation se focalisent sur une échelle locale car ils permettent aux navigateurs d'avoir une vision locale de leur environnement et que les systèmes de surveillance maritime se focalisent plutôt sur une échelle globale comme les territoires et les routes maritimes.

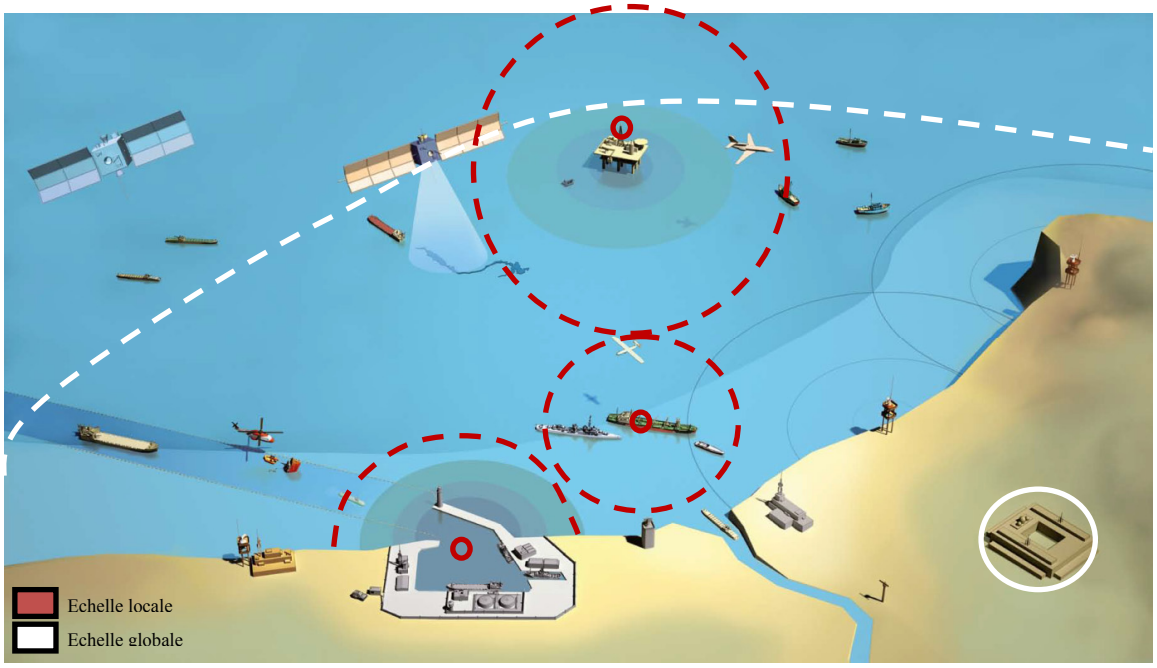


Figure 0-4 : Les différents échelles de surveillance maritimes (Thales 2013)[18].

L'échelle temporelle quant à elle permet de classer l'observation du trafic maritime en différentes fenêtres temporelles : une analyse *a posteriori*, temps-réel et anticipé (Figure 0-5). Les deux systèmes de surveillance, à savoir, les systèmes d'aide à la navigation et les systèmes de surveillance du trafic maritime affichent les dernières positions des navires pour permettre aux utilisateurs de ces systèmes d'analyser les comportements en temps-réel. Ces systèmes permettent d'identifier automatiquement des risques potentiels de collisions en se basant sur les cinématiques des navires (Etienne 2011).

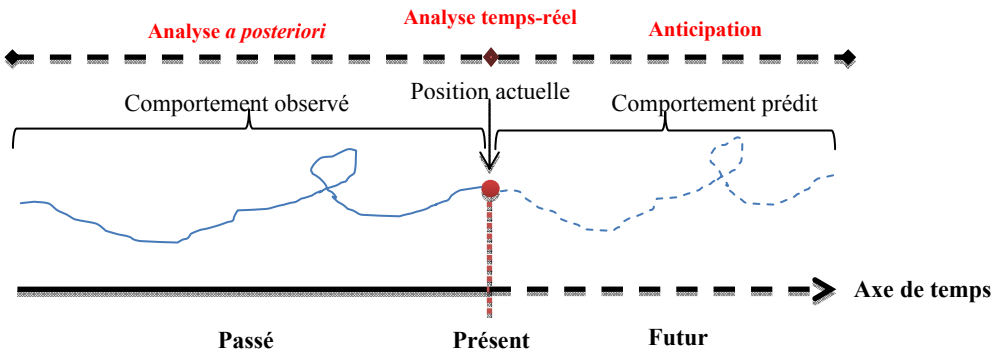


Figure 0-5 : Illustration des différentes fenêtres d'analyse des comportements de navires : a posteriori, temps-réel et anticipé.

[18] http://www.thalesgroup.com/Markets/Security/Documents/Maritime_Safety_and_Security/

Un risque maritime toujours aussi important

Bien que, des dispositifs réglementaires, organisationnels et technologiques soient déployés, les risques maritimes sont toujours aussi importants. Les statistiques recensées par les différents organismes internationaux montrent l'importance de ces risques. Le Bureau Maritime International (BMI) recense 445 actes de piraterie en 2010 avec une croissance de 8,5% par rapport à 2009 et 1181 marins pris en otage dans la même année (Deiss 2011). La piraterie coûte chaque année à l'Union Européenne 565 millions d'euros, principalement dus au réacheminement de marchandises pour éviter les zones à risque, les primes de risques et d'assurance kidnapping. Le Centre de documentation de recherche et d'expérimentations sur les pollutions accidentelles des eaux (Cedre), publie que 54 700 tonnes d'hydrocarbures et de substances dangereuses ont été déversées accidentellement en 2009 contre 7 500 tonnes en 2008 avec une augmentation importante de plus de 600%. Sans oublier les milliers d'accidents et d'incidents maritimes recensés par an dans le monde mettant en danger les personnes, l'environnement et l'économie. Selon l'Organisation des Nations Unies pour l'alimentation et l'agriculture (FAO), la pêche illégale représente 30% de la pêche totale dans certaines pêcheries importantes (FAO 2010) ce qui coûte plus d'un milliard d'euro par an à l'UE (Autran 2011). Selon la FAO, 19% des stocks de pêches qu'elle suit sont surexploités et 8% sont même épuisés[19]. L'organisation Environmental Justice Foundation (EJT) publie qu'en 2009, 16% des importations de l'UE proviennent d'une pêche illégale ce qui peut avoir des répercussions importantes sur le secteur maritime[20]. En ce qui concerne la lutte antidrogue, la facture payée chaque année par l'Europe s'élève à 18 milliards d'euros sachant que 70% de la cocaïne et 60% du cannabis sont saisis à bord de navires.

Pour réduire les risques, plusieurs pistes peuvent être envisagées. Parmi ces pistes, nous pouvons imaginer la proposition de nouvelles réglementations, une augmentation des moyens de contrôle et d'investigation et l'amélioration des systèmes de surveillance maritime. C'est cette dernière piste que nous privilégions dans notre travail de recherche.

[19] http://lci.tf1.fr/science/environnement/fao-stop-ou-encore-a-la-peche-illegale-5836838.html
[20] http://edition.cnn.com/2011/WORLD/africa/06/30/illegal.pirate.fishing/

Amélioration des systèmes de surveillance maritime

L'amélioration de la surveillance maritime peut concerner les infrastructures de détection de données ou les systèmes de traitement d'informations. Au niveau de la détection, de nouveaux capteurs plus performants sont proposés (FMCW[21], satellites, drones aériens, etc.) et en ce qui concerne les améliorations des systèmes de traitement d'informations, elles vont être détaillées ci-après.

L'une des pistes d'amélioration intéressante des systèmes de traitement d'informations est de les doter d'un module d'identification automatique de comportements à risques. Le comportement d'un objet mobile peut être défini comme un ou plusieurs mouvements dans une situation qui correspond à des facteurs et à des zones (zones à risque de piraterie, zones accidentogènes, etc.). Les facteurs sont des caractéristiques de l'objet (Type, dimension, etc.), de son contexte (panne, phare cassé, etc.) et de son environnement d'évolution (météorologie, océanographie, etc.). Le mouvement quant à lui est défini comme une évolution spatio-temporelle d'un mobile, un changement de positions pour les navires.

En effet, l'identification des comportements à risques par les opérateurs de surveillance est difficile et complexe à cause du nombre important de navires en déplacements, la multiplicité des risques et de leur caractère imprévisible. Les opérateurs se trouvent rapidement submergés par une quantité importante de cinématiques de navires qui rendent difficile l'identification de ces comportements. Chaque opérateur a jusqu'à 250 navires à surveiller (Etienne 2011) et selon la LIoyd's[22], 41 millions de positions AIS sont captées chaque jour pour 62 000 navires en mer.

Pour pouvoir automatiser cette identification, il faut au préalable disposer de modèles décrivant les comportements à risques. Des phases d'observation, d'analyse, de compréhension et de construction de connaissances sont nécessaires à cette modélisation. Ce travail de recherche s'inscrit dans cette perspective en proposant l'extraction de connaissances sur les mouvements et les situations potentiellement à risques, pour aider à la modélisation des comportements à risques. L'extraction de connaissances se base sur une approche inductive (partir des données pour les transformer en connaissances) et la

[21] Frequency Modulated Continuous Wave – Onde Continue Modulée en Fréquence
[22] Compagnie d'assurance.

découverte de régularités et irrégularités sous forme de motifs et de règles dans les historiques de données. Ces motifs et règles peuvent aider à l'analyse *a posteriori* des comportements à risques en identifiant par exemple des facteurs, zones et mouvements potentiellement à risques. Un motif désigne un modèle, une structure et une forme décrivant une organisation apparaissant dans les observations. Appelé aussi pattern (terme en anglais), le motif est une sorte de synthèse ou de résumé des ensembles d'observations (Etienne 2011) apparaissant d'une manière répétée dans le cas de motifs fréquents. Une règle est une expression de la forme A $\mathbb{R}$ B tel que A et B sont des ensembles de valeurs de variables et $\mathbb{R}$ une relation entre les deux ensembles. La relation peut être par exemple d'association ou de corrélation (Cf. section 2.2.1.1).

Dans ce travail de thèse, il est proposé une méthodologie pour l'extraction de connaissances sous forme de motifs et de règles décrivant des comportements à risques. Cette méthodologie est composée : d'un ensemble de méthodes de fouille de données ; d'une méthode d'acquisition et de préparation de l'espace de données à explorer ; et enfin, d'une méthode d'analyse et de validation de notre proposition à l'aide d'un atelier fondé sur cette méthodologie.

L'extraction de connaissances à partir de grands volumes de données demande de suivre un certain nombre d'étapes regroupées sous forme d'un processus d'Extraction de Connaissance à partir de Données (ECD). Une ECD (*Knowledge Discovery in Databases* (KDD) en anglais), regroupe l'ensemble des méthodes et des outils qui permettent de transformer les données volumineuses et hétérogènes en connaissances utiles à la prise de décision. Les méthodes d'ECD doivent être ordonnées dans un processus logique, d'extraction de données, sélection, préparation, exploration et diffusion des résultats aux consommateurs d'informations.

Ces méthodes sont issues de plusieurs domaines comme les bases de données, l'intelligence artificielle et les statistiques. La fouille de données (*data mining*) fait partie intégrante de l'ECD et représente l'étape d'exploration de données. La fouille de données peut être définie comme l'extraction non triviale de connaissances cachées et potentiellement utiles à partir de données (Frawley et al., 1992). Elle permet l'exploration automatique des données et la découverte de nouvelles connaissances non connues auparavant.

Problématique et objectifs de recherche

Notre travail de thèse a porté sur la fouille de données appliquée à la gestion des risques maritimes et plus particulièrement sur l'aide à la modélisation des comportements à risques de navires.

Aujourd'hui, énormément de données brutes sont produites sur le déplacement des navires et leur environnement de navigation. L'analyse de ces données présente un enjeu majeur pour extraire des connaissances sur les comportements de navires. L'exploration de ces données par des méthodes de fouille de données va générer automatiquement des motifs et des règles décrivant des comportements inhabituels, fréquents, périodiques et de groupements.

La problématique de notre travail est de savoir si les motifs (spatio-temporels inhabituels, fréquents, périodiques, groupements, etc.) et les règles (règles d'association, séquentielles, etc.) issus de la fouille de donnés peuvent décrire un comportement à risque. Si oui alors quelles méthodes et algorithmes de fouille de données utiliser pour les extraire ?

Selon l'échelle temporelle vue sur la Figure 0-5, la fouille de données se place dans la fenêtre d'analyse *a posteriori* en permettant l'exploration des comportements passés. L'objectif de cette exploration est d'aider à construire des modèles descriptifs et prédictifs des comportements. Les modèles descriptifs peuvent être utilisés pour l'analyse temps-réel des comportements de navires et les modèles prédictifs pour l'anticipation des comportements.

Méthodologie

La problématique posée dans la section précédente est identifiée à partir d'une recherche bibliographique et d'une réflexion guidée par l'implication de notre centre de recherche dans plusieurs projets de surveillance maritime comme ScanMaris (Morel et al., 2008; Morel et al., 2010), TaMaris (Morel et al., 2011), SisMaris (Morel 2009), I2C (Morel & Broussolle 2011)) et dans différents travaux de recherche (Vandecasteele 2012)(Idiri & Napoli 2012b)(Chaze et al., 2012)(Vatin & Napoli 2013a).

Il est intéressant de noter que la plupart des approches actuelles d'acquisition de connaissances sur les comportements de navires sont basées sur l'expertise humaine ce qui présente quelques limites qu'il est intéressant de combler. Ces limites sont exposées dans la section 1.7 du chapitre 1.

Afin de combler ces limites, une méthodologie d'extraction de connaissances sur les comportements à risques de navires est proposée. La méthodologie se compose de plusieurs phases :

- **Analyse des besoins sur la découverte de connaissances décrivant des comportements à risques**,

  Dans ce travail de recherche, les besoins d'extraction de comportements potentiellement à risques de navires sont issus d'une revue de littérature (Vandecasteele & Napoli 2012) (Etienne 2011) (Morel et al., 2010)(Marven et al., 2007).
- **Identification de méthodes de fouille de données permettant l'extraction de ces comportements à risques de navires**,

  Dans cette phase, un ensemble de méthodes de fouille de données sont identifiées et testées pour l'extraction de connaissances décrivant des comportements de navires potentiellement à risques.
- **Acquisition et préparation de données à explorer**,

  Dans cette phase, nous devons répondre à deux questions : quelles données doit-on acquérir ? Comment les préparer et les exploiter pour en extraire des connaissances sur les comportements à risques ?
- **Analyse et validation des méthodes de fouille de données pour l'extraction de comportements à risques de navires à l'aide d'un atelier fondé sur cette méthodologie**

  Dans cette phase, il est décrit comment extraire des connaissances sur les comportements de navires, les analyser et comment interpréter les comportements potentiellement à risques. La validation des méthodes de fouille de données identifiées passe par la conception et le développement d'un atelier d'analyse de comportements à risques fondée sur cette

méthodologie. L'atelier va offrir un environnement destiné à l'analyse et à l'aide à la modélisation des comportements à risque.

Nous présentons ci-après (Figure 0-6), un diagramme représentant les étapes générales de la méthode d'avancement suivie dans cette thèse. Les lectures bibliographiques effectuées ont été organisées en plusieurs thèmes à savoir, la surveillance maritime, la fouille de données et l'analyse de comportements.

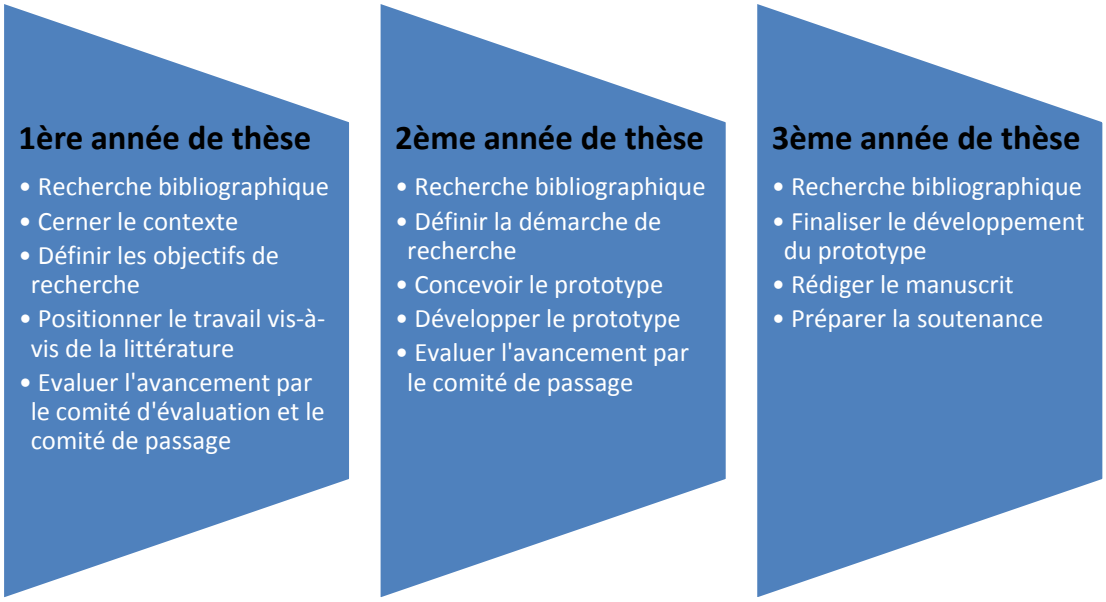


Figure 0-6 : Diagramme représentatif de la méthode d'avancement de la thèse.

Périmètre de notre proposition

L'exploration des comportements de navires est basée sur les données AIS qui présentent des limites concernant la portée du signal radio et le nombre de navires qui ont obligation à s'équiper de ce transpondeur. Les infrastructures d'acquisition AIS actuelles ne permettent pas l'identification des navires qui sont loin des côtes, les petites embarcations et les navires non coopérants (qui ne sont pas équipés de transpondeur). Cette incomplétude des données empêche la découverte des comportements à risques de ces navires.

Dans ce travail, nous ne proposons pas une nouvelle méthode de construction de connaissances mais une nouvelle approche combinant plusieurs méthodes de fouille de

données. Ces méthodes ont été étudiées, sélectionnées et intégrées afin de démontrer la validité de la méthodologie. La plus-value de ce travail est la découverte automatique de comportements à risques de navires.

Le dernier périmètre concerne le cadre applicatif de l'étude. Nous avons choisi d'appliquer notre méthodologie à une problématique maritime. Il est alors nécessaire d'écarter les navires naviguant dans les fleuves.

Structure de ce mémoire

Ce mémoire se compose de quatre chapitres. Le chapitre 1 concerne l'analyse de comportements dans le domaine de la surveillance maritime ; le chapitre 2 porte sur la contribution de la fouille de données à l'analyse de comportements ; le chapitre 3 est consacré à la conception et au développement d'un environnement d'extraction de connaissances sur les comportements à risques et enfin ; le chapitre 4 présente quelques exemples d'extraction de comportements potentiellement à risques pour valider notre méthodologie. Nous terminerons le mémoire par une conclusion qui pose le bilan de ce travail, met en avant les principales contributions et présente de multiples perspectives approfondies.

Chapitre 1 : L’analyse de comportements dans le domaine de la surveillance maritime

1.1. Introduction

Les progrès technologiques en systèmes de localisation (GPS, RFID, etc.), de télétransmission (VHF, satellite, GSM, etc.), en systèmes embarqués et leur faible coût de production a permis leur déploiement à une large échelle. Enormément de données sur le déplacement d'objets sont acquises par le biais de ces technologies et utilisées dans diverses applications comme le suivi de foules de piétons (Giannotti et al., 2011) (Buard & Christophe 2012), le suivi des animaux (Buard & Brasebin 2011), la gestion du trafic routier, aérien et la surveillance maritime (Etienne 2011). Des bases de données de suivi d'objets mobiles sont mêmes mises en libre utilisation sur internet comme les bases de données AISHUB[23] de suivi de navires.

La plupart du temps ces données sont utilisées pour des besoins temps-réel. L'analyse *a posteriori* des historiques de données peut présenter des perspectives très intéressantes pour l'analyse des comportements et la compréhension des mouvements, des situations et de leur interconnexion.

Dans ce chapitre, nous présentons l'analyse de comportements d'objets mobiles et l'intérêt produit pour plusieurs domaines d'application. Nous décrivons ensuite, les approches d'acquisition et de construction de connaissances utilisées pour la modélisation des comportements. Puis, nous allons nous focaliser sur le domaine maritime et présenter les principales méthodologies de modélisation qui ont été utilisées pour l'analyse et la modélisation des comportements de navires. Ces travaux ont montré l'intérêt accru des méthodes d'analyse de comportement pour la surveillance maritime. Enfin, nous allons présenter une synthèse récapitulative de ces méthodologies (description, avantages et limites).

1.2. Définition d'un comportement

Le comportement est défini dans le dictionnaire LAROUSSE[24] comme étant une action, une réaction, un fonctionnement et une évolution spatio-temporelle dans certaines situations d'un objet. Cet objet peut être un véhicule, un navire ou toute autre chose. Dans la suite de ce travail, nous considérons un comportement comme un ensemble de

[23] http://www.aishub.net/
[24] http://www.larousse.fr/dictionnaires/francais/comportement/17728

mouvements dans une situation. Le mouvement est un changement de position et la situation est une combinaison de caractéristiques de l'objet, de son contexte et de son environnement d'évolution spatio-temporel. Le mouvement est influencé par la situation qui peut être liée à des facteurs internes à l'objet (objectif, panne, etc.) ou des facteurs externes (mouvements des voisins, météorologie, etc.) (Le Pors et al., 2009) dans (Etienne 2011). Un navire qui change son cap par exemple pour éviter de rentrer en collision avec un autre navire est un mouvement influencé par l'environnement dans lequel il évolue.

1.3. L'analyse de comportements

L'une des problématiques de l'étude des comportements d'objets mobiles est de comprendre les mouvements, les interactions entre les objets, leur environnement et comment il est possible d'interpréter ces comportements à partir des propriétés quantitatives et qualitatives observées des déplacements. L'étude des déplacements a pour objectif de faciliter la compréhension des causes, des mécanismes, des patrons spatio-temporels du mouvement et leur rôle dans l'évolution du système (Nathan et al., 2008). Les régularités et les irrégularités dans les mouvements peuvent être décrites par des motifs (patrons ou modèles). Un motif peut être considéré comme la synthèse des mouvements, une description des comportements et un modèle de prédiction.

Les positions des mouvements réels sont représentées par des coordonnées affichées sur une cartographie. La jonction de positions d'un même objet permet de représenter l'évolution de l'objet par une trajectoire qui caractérise son déplacement. Une trajectoire est donc une suite de positions ordonnées dans le temps.

L'analyse de comportements des objets à partir de ces trajectoires ouvre des perspectives intéressantes dont la compréhension, la description et la prédiction de la mobilité. Cette capacité de recueille et d'analyse de quantités massives de données de déplacement ont transformé plusieurs domaines comme la biologie et les sciences sociales en permettant d'interpréter les comportements (Lazer et al., 2009). Qu'en est-il du domaine de la surveillance maritime ?

La problématique d'analyse automatique des comportements de navires a sollicité un fort intérêt au niveau des organismes et centres de recherche traitant de la question de la sûreté maritime. Le projet Predictive Analysis for Naval Deployment Activities (PANDA) (Darpa 2005) du ministère de la défense américain a pour objectif d'évaluer automatiquement les comportements de tous les grands navires et pas seulement ceux qui font l'objet d'un suivi (*Vessel Of Interest*), afin de déterminer quels sont ceux qui s'écartent de leur comportement normal et attendu[25]. L'objectif est d'indiquer les menaces relatives à la sûreté maritimes par analyse automatique de comportements de navires. Ce projet est considéré comme initiateur et a inspiré de nombreux travaux dont le projet SARGOS (Chaze et al., 2012) pour la protection de plateformes pétrolières, le projet SECMAR pour la surveillance des ports, le projet ScanMaris (Morel et al., 2008; Morel et al., 2010), TaMaris (Morel et al., 2011), SisMaris (Morel 2009) pour l'analyse du trafic maritime et les projets I2C et Perseus pour l'analyse du trafic maritime et l'interopérabilité des systèmes de surveillance européens.

La recherche et développement (R&D) pour la Défense Canadienne, s'est intéressé aussi à cette question en proposant l'analyse visuelle des données maritimes (Gouin et al., 2011; Lavigne & Gouin 2011). D'autres travaux académiques comme les travaux de N. Willems et M. Riverio ont proposé des méthodes d'analyse visuelle des données maritimes (Willems et al., 2009; Willems 2011; Willems et al., 2011; Riveiro et al., 2008; Riveiro & Goran Falkman 2009; Riveiro & Göran Falkman 2011).

Afin de détecter automatiquement les comportements anormaux, il est possible de modéliser ce qui est anormal pour identifier les comportements qui suivent ce modèle ou de modéliser ce qui est normal pour identifier les comportements qui s'écartent de cette normalité. Les deux approches sont utilisées pour la construction de modèles de comportements (Kai-Lin et al., 2013). Prenons l'exemple d'une étude de mouvements d'utilisateurs de parking à partir d'enregistrements vidéos (Figure 1-1). La découverte d'un mouvement inhabituel ou suspect faisant des déplacements aléatoires entre les véhicules du parking peut décrire un comportement d'un voleur qui cherche quelle voiture voler (partie B de la Figure 1-1). La partie A décrit un comportement normal et la partie B, un comportement aléatoire qui s'écarte de la normalité.

[25] https://www.fbo.gov/spg/ODA/DARPA/CMO/BAA05-44/listing.html

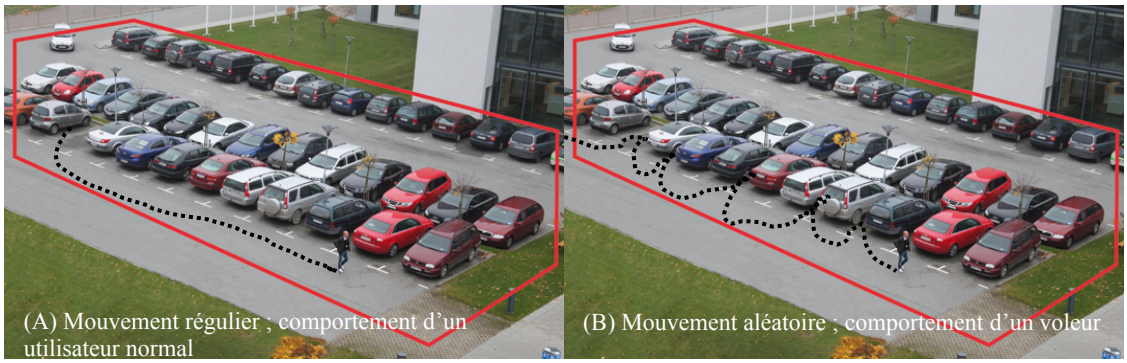


Figure 1-1 : Identification d'un comportement suspect à partir de l'analyse du déplacements d'une personne sur une vidéo de surveillance (Han, 2010) [26].

Un autre exemple présente ci-dessous une trajectoire inhabituelle d'un navire qui fait un demi-tour et revient s'arrêter devant la côte. Cette trajectoire peut correspondre à une dérive d'un navire comme le montre la Figure 1-2 représentant la dérive puis l'échouement du Costa Concordia sur les côtes italiennes en janvier 2012.

Figure 1-2 : Trajectoire du Costa Concordia au moment de l'échouement.

Le mouvement à lui seul n'est pas toujours suffisant pour analyser les comportements, il est alors nécessaire d'étudier la situation dans laquelle évolue le mouvement. L'analyse de situations peut aider à mieux qualifier un comportement.

[26] Exemple inespéré de la présentation de J. Han à DASFAA 2010, slide 97, les images sont récupérées de http://www.axis.com/fr/academy/10_reasons/camera_intelligence.htm

1.4. Approches d'acquisition et de construction de connaissances pour la modélisation de comportements

Dans la construction de connaissances, deux approches peuvent être utilisées (Roddick & Lees 2009) :

- Approche déductive : modélisation du monde réel avec des modèles mathématiques dont la prévision ou la description des modèles est calculée,
- Approche inductive : correspondance de modèles dont la prévision est faite par rapport aux observations passées :
  - A partir de l'expérience (*top-down*) dont l'observation et le raisonnement est fait par l'humain,
  - A partir de l'analyse des bases de données[27] (*bottom-up*)
    - Analyse statistiques,
    - Analyse visuelle de données,
    - Fouille de données.

La déduction permet de déduire des conséquences observables à partir d'hypothèses générales (prémisses) (Martin 2012). Contrairement à la déduction, l'induction produit des prémisses et offre la possibilité de générer de nouvelles connaissances. Cette approche de raisonnement passe d'observations particulières à des hypothèses générales en faisant des restrictions sur un espace d'hypothèses jusqu'à ce qu'une description restrictive de cet espace puisse être formée (Roddick & Lees 2009).

L'induction part de l'idée que « *la répétition d'un événement augmente la probabilité de le voir se reproduire* ». La répétition d'un événement n'implique pas forcement sa reproduction. Par conséquent, cette approche privilégie l'observation, l'analyse et l'expérimentation pour tirer des conclusions générales (Martin 2012). La validation des connaissances issues de l'induction est donc primordiale pour éviter d'avoir des connaissances inutiles ou erronées. C'est pourquoi, nous considérons que cette approche est à privilégier pour la découverte de connaissances.

[27] Une base de données est un ensemble de données organisées dans des structures pour faciliter la gestion des données (ajout, suppression, mise à jour, interrogation, etc.) et leur utilisation par des applications informatiques. (Gardarin 2011).

La Figure 1-3 résume bien les approches d'acquisition et de construction de connaissances à savoir, la déduction et l'induction. La fouille de données appartient à la correspondance de modèles (Pattern Matching) qui correspond à l'approche inductive.

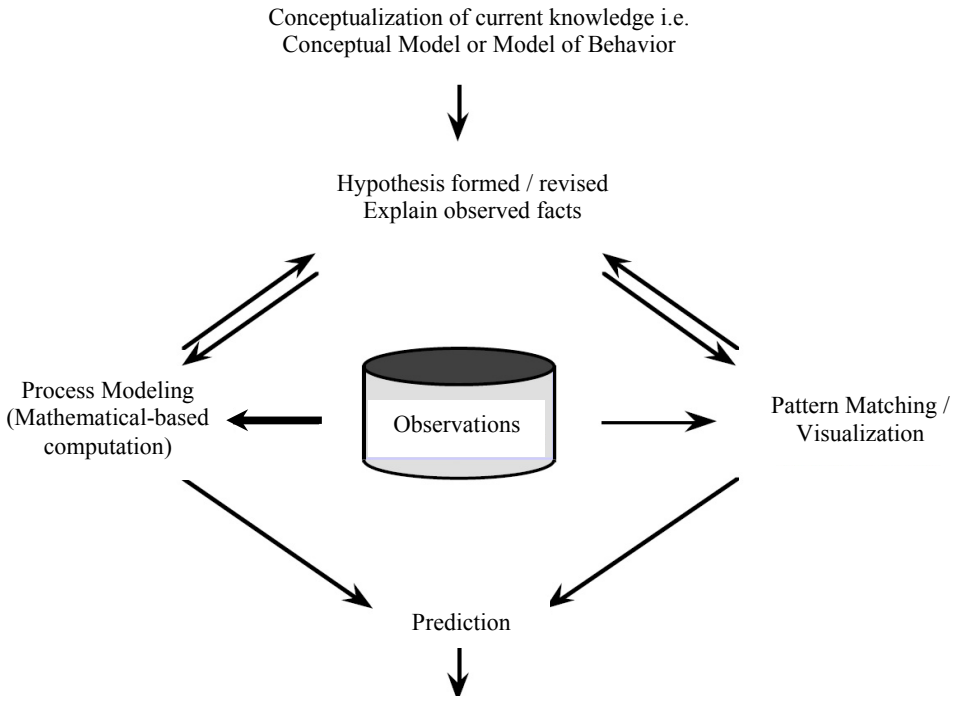


Figure 1-3 : Les approches d'acquisition de connaissances
(Roddick & Lees 2009).

Dans l'approche inductive, deux sous approches sont utilisées : une approche *top-down* qui permet la découverte de connaissances à partir du raisonnement humain sur des observations et une approche *bottom-up* guidée par des explorations automatiques de bases de données d'observations.

Nous les avons appelés ainsi car dans l'approche *top-down,* le mécanisme d'apprentissage commence à partir de l'expertise alors que dans l'approche *bottom-up,* il commence à partir d'observations stockées dans des bases de données.

1.4.1. Approche *top-down*

Dans cette approche basée sur l'expertise, la méthode de brainstorming est souvent utilisée ((Roy 2008) ; ((Nilsson et al., 2008) dans (Laere & Nilsson 2009))). Le *brainstorming* consiste à réunir des experts pour l'acquisition de connaissances sur les

comportements habituels, inhabituels ou suspects. Ces méthodes sont compliquées à mettre en œuvre et couteuses ; les scénarios en sortie dépendent beaucoup de l'expérience personnelle de chaque expert. De plus, elles ne permettent pas la découverte de connaissances nouvelles.

1.4.2. Approche *bottom-up*

Cette approche d'apprentissage basée sur les données, utilise plusieurs méthodologies pour l'acquisition de connaissances. Parmi ces méthodologies nous pouvons citer les analyses statistiques, la visualisation, l'apprentissage automatique et la fouille de données.

L'intérêt de la fouille de données est qu'elle rassemble plusieurs techniques statistiques, d'apprentissage machine, algorithmiques, etc. pour exploiter les avantages de chacune de ces techniques. De plus, les avancées informatiques en calcul et en stockage permettent l'analyse rapide de grands volumes de données.

1.5. Méthodologies d'analyse de comportements de navires

Le domaine d'application qui a été privilégié dans cette thèse est l'étude des comportements de navire. Pour l'analyse de ces comportements, différentes méthodologies ont été proposées. Dans les sous sections suivantes, nous allons présenter l'analyse statistique, l'analyse visuelle et la fouille de données.

1.5.1. Analyse statistique

Les statistiques peuvent être définies comme un ensemble de techniques et méthodes permettant le traitement et l'interprétation des données. Selon J. Tukey (Tukey 1980), il existe deux approches d'analyse statistiques, une analyse exploratoire et une autre confirmatoire. L'analyse exploratoire part des données qu'elle permet d'analyser sous différentes facettes pour mettre en évidence des structures cachées par le volume important des données et aider ainsi à construire des modèles (Ladiray 1997). Le volume de données analysé peut atteindre plusieurs dizaines de milliers d'individus et des dizaines de variables. Cette analyse peut précéder une analyse confirmatoire pour mettre en exergue les propriétés qualitatives, quantitatives des données et poser des hypothèses

plausibles. Pour infirmer ou confirmer une hypothèse, c'est l'analyse confirmatoire qui intervient. Cette analyse est faite sur un échantillon de données représentatif de la population globale puis les informations résultantes sont inférées à la population entière avec un certain degré de confiance. Cette inférence peut être vue comme une extrapolation de nouvelles informations à partir de celles connues déjà. L'analyse exploratoire et confirmatoire peuvent et doivent passer d'une à l'autre car elles sont complémentaires.

Dans (Etienne et al., 2010), les auteurs ont proposé une analyse statistique appliquée à un historique de déplacement de navires qui est décrit par des trajectoires spatio-temporelles. Leur proposition se fond sur une ingénieuse extension de la boite à moustaches[28] à l'analyse spatio-temporelle. Une médiane spatiale et temporelle sont calculées dans l'objectif de les utiliser dans la définition du couloir spatio-temporel (Figure 1-4). Le couloir spatial et temporel est défini d'une manière à ce qu'ils incluent 90% des positions les plus proches de la médiane. C'est le 9ème décile[29] qui a été pris comme séparateur entre les 90% des positions et les 10% restantes. Le couloir spatio-temporel résultant de l'analyse a été utilisé comme motif pour détecter les comportements inhabituels à partir de trajectoires. Les comportements inhabituels sont découverts par intersection d'une position avec ce motif pour la qualifier à un instant t de position en retard, à l'heure, en avance, sur la route ou en dehors du couloir spatio-temporel.

[28] Appelée en anglais Box Plot, c'est une représentation graphique en boite permettant de représenter schématiquement la distribution d'une série statistique.

[29] Est une statistique descriptive contenant neuf valeurs séparant un ensemble d'individus en parts égales en effectif.

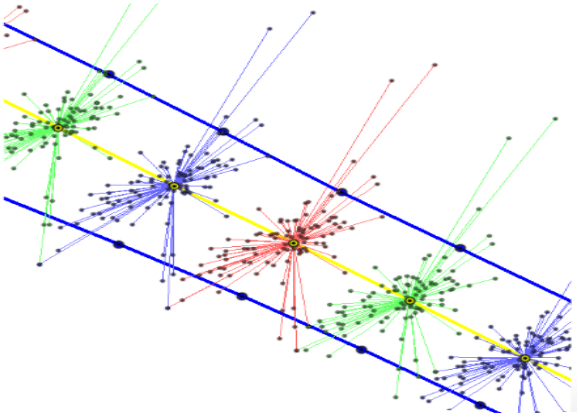


Figure 1-4 : Définition de positions médiane, la trajectoire médiane et du couloir spatio-temporel pour l'identification des comportements inhabituels de navires (Etienne et al., 2010).

1.5.2. Analyse visuelle

L'analyse de comportements de navires par exploration visuelle des données utilise des techniques de géo-visualisation pour présenter les informations d'une manière à pouvoir en extraire du sens. Willems (Willems et al., 2011) par exemple propose un algorithme de visualisation des trajectoires de navires basé sur la densité : plus un navire navigue rapidement, moins il laisse de traces. Ces traces représentent des informations sur les routes maritimes, des zones d'ancrage, les lenteurs et déplacements rapides et bien d'autres informations (Figure 1-5).

Dans ce genre d'approche c'est à l'utilisateur d'interagir avec les représentations visuelles pour identifier les anomalies et construire des connaissances. Les techniques de visualisation vont aider à améliorer les capacités de perception, de compréhension et de raisonnement de l'être humain (Riveiro & Goran Falkman 2009).

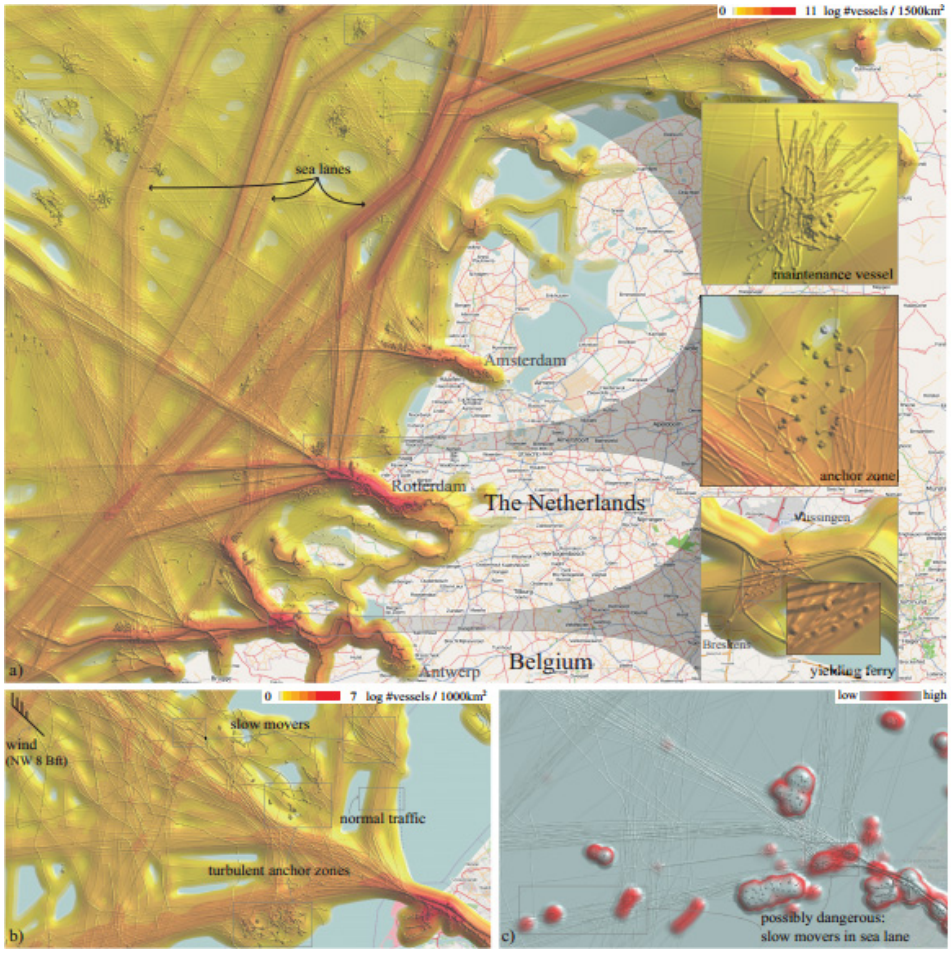


Figure 1-5 : Analyse des densités de déplacement de navires (Willems et al., 2009).

Dans (Riveiro & Goran Falkman 2009), les auteurs ont utilisé un graphique de fréquence de vitesses pour extraire par visualisation des comportements inhabituels de navires par rapport à leur vitesse (profil de navires lents et trop rapides). Dans un autre travail (Etienne et al., 2011), les auteurs utilisent un indice de similarité pour détecter par géo-visualisation les trajectoires outliers qui ne suivent pas la même progression temporelle et spatiale que le groupe de trajectoires.

L'avantage de l'analyse visuelle et géo-visuelle est lié à l'intégration de la dimension humaine dans l'exploration de données. Dans cette analyse, les utilisateurs interagissent avec différentes formes de visualisation dans l'objectif d'extraire des connaissances. Cette participation des utilisateurs dans la construction des connaissances est intéressante car ils connaissent le domaine d'application.

Les inconvénients de cette méthode d'analyse de comportements sont la complexité de certaines visualisations et le temps requis pour l'affichage et l'exploration des données. Prenons l'exemple d'une forme de géo-visualisation appelée Trajectory Wall (G. Andrienko et al., 2014) qui montre la complexité de certaines représentations pour des utilisateurs non habitués à ce genre de représentation (Figure 1-6). Les objets rapides sont représentés en vert et les objets lents en rouge. La représentation circulaire quant à elle, permet de voir la moyenne des intervalles de vitesses par tranche horaire.

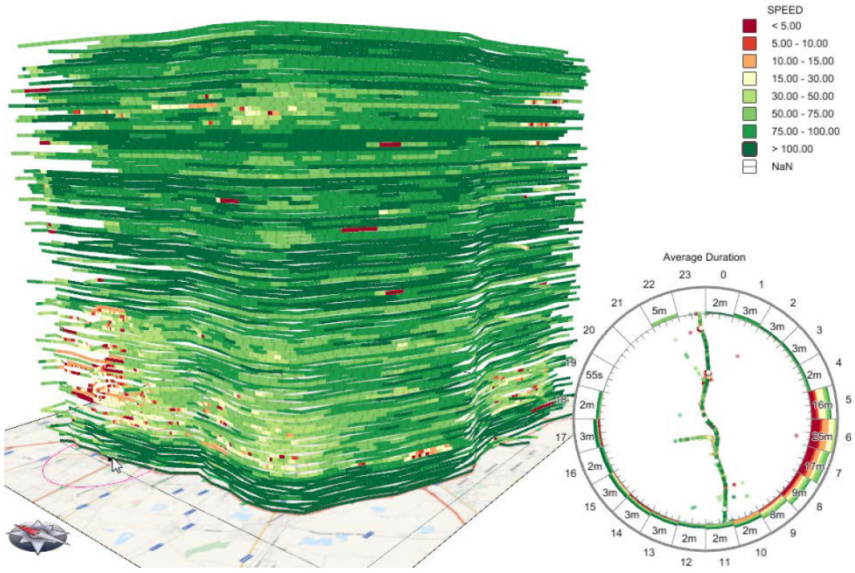


Figure 1-6 : Géo-visualisation de comportements d'objets mobiles par distribution d'intervalles de vitesses dans le temps
(G. Andrienko et al., 2014).

Généralement le problème ne réside pas dans la complexité des visualisations mais plutôt dans l'inadaptation de ces visualisations aux profils des utilisateurs et à l'analyse de comportement effectuée. C'est pour cela que Vatin (Vatin & Napoli 2013a)(Vatin & Napoli 2013b), propose dans son travail de recherche, une approche permettant d'adapter automatiquement la visualisation selon le profil de l'utilisateur et l'analyse de comportements qu'il souhaite effectuer.

1.5.3. Analyse par fouille de données

Par analogie à la recherche des pépites d'or dans un gisement, la fouille de données vise à extraire des informations cachées par analyse globale et à découvrir des modèles,

appelés motifs, difficiles à percevoir directement du fait du volume important des données, du nombre de variables à considérer et enfin du fait qu'il y ait des hypothèses imprévisibles (Gardarin 2011).

La fouille de données peut être vue comme un générateur automatique d'hypothèses pour examen et validation dans le but de transformer les données en connaissances. Plusieurs définitions de la fouille de données existent. Parmi ces définitions, nous pouvons citer celle de Frawley (Frawley et al., 1992) qui présentent la fouille de données comme une extraction non triviale de connaissances implicites et potentiellement utiles à partir des données. Parsaye (Parsaye 1995), le définit comme un processus d'aide à la décision où les utilisateurs cherchent des modèles d'interprétation dans les données. Une autre définition plus anecdotique est celle de Chorafas : la fouille de données consiste à torturer les données jusqu'à ce qu'elles avouent.

Il y a dans la littérature des travaux initiateurs dans l'utilisation des méthodes de fouille de données au problème de surveillance maritime (LeBlanc & Rucks 1996) (Torun & Düzgün 2006) (Marven et al., 2007) (Etienne et al., 2010). Nous allons décrire quelques-uns de ces travaux.

1.5.3.1. Analyse de situations par clustering d'événements

Le Blanc et Rucks sont parmi les pionniers à avoir utilisé la fouille de données au domaine maritime (LeBlanc & Rucks 1996). Ils ont appliqué une méthode de clustering des plus proches voisins KNN (Anderberg 1973) sur un échantillon de 936 cas d'accidents qui se sont produits sur le fleuve du Mississippi. Cette méthode de clustering avait permis d'identifier quatre clusters nommés zones dangereuses, mauvaises conditions de navigation, accidents pouvant être évités, accidents qui n'auraient pas dû arriver. Les zones dangereuses regroupent les accidents qui se sont produits dans des parties dangereuses du fleuve.

Après avoir identifié ces clusters une analyse discriminante est réalisée pour déterminer les valeurs d'attributs qui séparent le mieux les cas d'accidents dans les différents clusters. Cette analyse discriminante a comme objectif de prédire le groupe de nouveaux cas d'accidents et de montrer l'intérêt de l'utilisation des systèmes de suivi de

navires. Les accidents classés comme pouvant être évités sont caractérisés par la non utilisation de ces systèmes.

L'un des inconvénients de cette méthode est le fait qu'elle construit les clusters en connaissant au préalable les zones dangereuses, comme les zones avec des rochers affleurant, les passages étroits et les zones à forte densité de trafic. La vocation de ce travail est d'utiliser les résultats obtenus pour classer les nouveaux accidents selon les quatre classes et à montrer ainsi l'intérêt d'utiliser un système de suivi de navires.

Torun et de son équipe (Torun & Düzgün 2006) ont aussi utilisé la fouille de données dans le domaine de la sécurité maritime. L'existence de points étroits dans le détroit[30] d'Istanbul, le fait qu'il soit rocheux, qu'il contienne des virages et des courants, etc., augmente les risques du transport maritime dans cette zone. Cela a motivé l'équipe de Torun à proposer un modèle linéaire de vulnérabilité[31] pour calculer plus précisément les risques en se basant sur les données de danger. Le danger peut être défini comme étant la menace qui pèse sur la sécurité, la sûreté, les personnes, l'activité et les biens. La particularité du danger est qu'il ne dépend pas de l'existence d'objets vulnérables mais existe indépendamment.

Torun et son équipe utilisent les techniques de fouille de données spatiales pour évaluer la vulnérabilité des personnes et des zones par rapport au transport de pétrole et de gaz dans le détroit d'Istanbul. Les techniques utilisées sont le clustering spatial avec les algorithmes K-means et ISODATA, l'autocorrélation, les hots spots et l'analyse de densité.

Le résultat de ce travail est présenté sur la Figure 1-7 où la partie (a) identifie les clusters de distribution d'accidents obtenus par l'indice de Moran[32]. Les distributions ayant un nombre d'accidents plus élevé que la majorité ou celles à proximité ont une couleur noir foncé. La partie (c) est le résultat de la superposition des zones de risques d'accidents (a) avec la vulnérabilité des personnes (b) décrivant les endroits pouvant avoir des conséquences graves en cas d'accident à proximité.

[30] Passage maritime naturel
[31] Conséquences de réalisation d'un risque sur les objets exposés.
[32] Mesure statistique d'autocorrélation proposée par Patrick Alfred Pierce Moran.

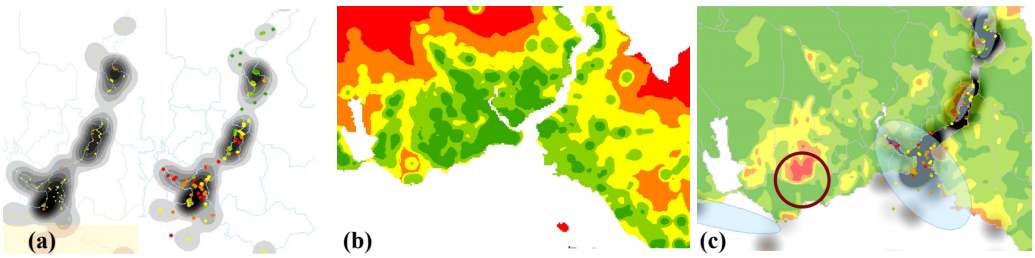


Figure 1-7 : (a) Zones à risques d'accidents, (b) Vulnérabilité des personnes, (c) Chevauchement de (a) et (b) (Torun & Düzgün 2006).

Dans un article de C. Marven (Marven et al., 2007), les auteurs ont montré l'intérêt des méthodes exploratoires à l'analyse des risques maritimes. Ils ont appliqué des méthodes d'analyse spatiale pour aider les gardes côtes canadiens à planifier leur recherche, prendre des décisions et améliorer le sauvetage en mer. Pour cela ils se sont basés sur le clustering spatial pour identifier et visualiser les concentrations d'incidents et d'accidents maritimes.

Les auteurs de l'article montrent l'opportunité d'appliquer au domaine maritime, des méthodes d'analyse spatiale utilisées en épidémiologie et criminologie. Parmi ces méthodes, on trouve Spatial and Temporal Analysis of Crime (STAC[33]) et Nearest Neighbour Hierarchical Cluster Analysis (NHH).

Concernant les deux dernières méthodes, ce ne sont pas des limites qui sont dressées mais une perspective. Les résultats des méthodes auraient pu être utilisés pour faire une étude de causalité des accidents, afin de comprendre les causes des concentrations anormales. La matérialisation de ces concentrations d'accidents sous forme de zones dans une base de données pourrait alerter en temps-réel des navires fréquentant ces zones.

1.5.3.2. Analyse du comportement par clustering de trajectoires

Dans (Etienne et al., 2010), les auteurs ont proposé une méthode d'extraction de trajectoires homogènes appelée Groupe Homogène de Trajectoires (GHT) qui donne de meilleures performances de calcul que la méthode T-Clustering, proposée par Andrienko

[33] http://www.icjia.state.il.us/public/index.cfm?metaSection=data&metaPage=stacfacts

(G. Andrienko et al., 2009) et intégrée dans M-Atlas (voir section 2.3.2). Concernant les résultats obtenus par les deux méthodes, ils sont presque identiques (Figure 1-8). Cette extraction de trajectoires ayant des mouvements similaires est basée sur deux notions différentes, à savoir les graphes d'intérêt et la similarité entre trajectoires (Etienne et al., 2008)(Etienne et al., 2009).

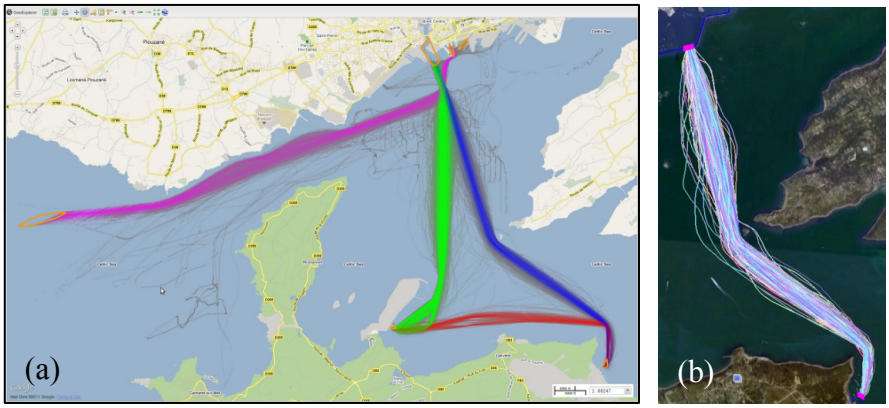


Figure 1-8 : Extraction de groupes de trajectoires-(a) par T-Clustering et (b) par GHT (Etienne et al., 2010).

La méthode de fouille de données proposée par Etienne (Etienne et al., 2010) permet d'extraire des routes types et des couloirs spatio-temporels pour des itinéraires donnés. La route type est construite en se basant sur le calcul de médianes à chaque ensemble de positions d'un GHT.

Le point fort de cette méthode est sa rapidité de calcul. Elle utilise des critères de sélection qui permettent d'optimiser les temps de réponses. Son point faible est dans le nombre important d'étapes de préparation des données au préalable de l'analyse spatio-temporelle. En effet, pour traiter et nettoyer les trajectoires, il faut suivre les étapes suivantes :

- Définir pour chaque trajectoire une position de départ et d'arrivée,
- Couper les trajectoires après un temps estimé à un trajet entre la position de départ et d'arrivée, pour ne pas avoir des allers-retours,
- Filtrer le groupe de trajectoires homogènes par l'algorithme Douglas-Peucker (2.2.3.1.4) pour gommer les imprécisions et les aberrations,

- Projeter les positions de départ sur une ligne de biais pour recaler la dimension spatiale (Figure 1-9),

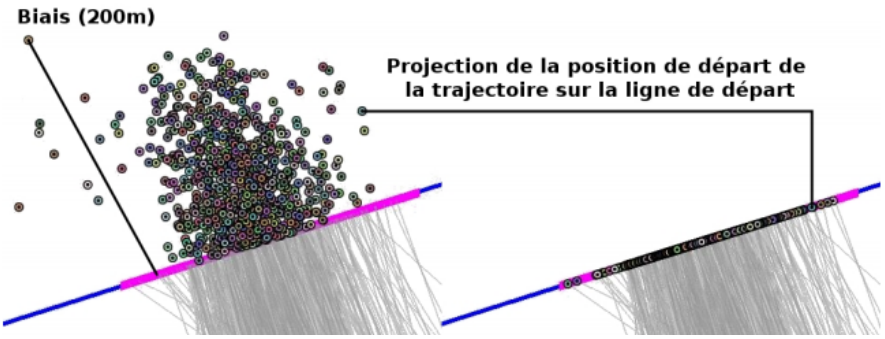


Figure 1-9 : Recalage spatial dans la méthode GHT (Etienne et al., 2010).

- Ré-échantillonner les positions des trajectoires homogènes pour avoir le même pas d'échantillonnage spatial. Par exemple, mettre une position chaque 100 mètres,

- Normaliser la dimension temporelle pour travailler sur des temps relatifs (temps écoulés depuis la position de départ).

Ces travaux montrent bien l'intérêt de l'utilisation de la fouille de données au domaine de la surveillance maritime : analyse de comportements inhabituels (navire en retard, en avance, en dehors de la route habituelle, etc.), découverte de routes types, de zones vulnérables, de concentrations d'accidents et la classification automatique de ces accidents.

1.6. Méthodologies de modélisation de comportements de navires

Différentes méthodologies de modélisation ont été utilisées pour l'analyse de comportements de navires. Dans les sous sections suivantes, nous allons décrire les principales méthodologies à savoir la modélisation par règles d'inférences, la modélisation ontologique et par classifieur Bayésien.

1.6.1. Modélisation par règles d'inférence

Le raisonnement automatique est un sous-champ de l'Intelligence Artificielle[34]. Il permet de simuler le raisonnement humain sur une machine pour déduire de nouvelles connaissances à partir de flux d'événements en entrée (faits avérés) et des connaissances mémorisées au préalable. Nous distinguons les faits issus d'événements en entrée du système de raisonnement que nous appelons *faits avérés* et les faits qui sont déduits par le système que nous appelons *faits inférés*.

Les connaissances dans un système de raisonnement sont souvent codées sous forme de règles (généralisation d'exemples) ou de cas (exemples). Une règle est de la forme « si *Antécédent* alors *Conséquent* » telle que, « *Antécédent* » et « *Conséquent* » sont des expressions de conjonction de disjonctions des occurrences d'objets. Une connaissance permet de mettre en relation des informations connues (« *Antécédent* ») et des informations qu'on cherche à déduire (« *Conséquent* ») ou des actions que l'on veut exécuter comme : si une condition est vérifiée alors déclencher une alerte (Jones 2007). Un cas est une description d'un problème avec sa solution associée. Ce paradigme enregistre des cas sources résolus dans une base de cas pour résoudre de nouveaux problèmes appelés cas cibles. Suivant la formulation de la connaissance, plusieurs types de raisonnement peuvent être appliqués. Les plus utilisés sont le raisonnement déductif et le raisonnement analogique. Dans le déductif, les valeurs en sortie sont déduites à partir des valeurs en entrée et ; dans le raisonnement analogique, le nouveau problème est ramené à un problème dont la solution est connue. La solution connue est donc adaptée au nouveau problème.

Un système de raisonnement à base de règles (RAPR) est choisi la plupart du temps par rapport au raisonnement à base de cas (RAPC) pour plusieurs raisons : tout d'abord, il est facile à comprendre car l'humain raisonne souvent sous forme de règles (si conditions alors actions). Puis, il permet la modularité sous forme de règles de connaissances. Un problème complexe peut être décomposé en règles simples. De plus, le raisonnement se fait par déduction et non par analogie. Le raisonnement par analogie peut aboutir parfois à des conclusions erronées (Lieber 2001).

[34] Domaine de recherche visant à rendre les machines intelligentes (raisonnement, perception, reconnaissance, etc.).

	RAPC	RAPR
Connaissance	Cas	Génération de cas
Modularité	problème	règle
Résolution des problèmes	Adaptation de cas	Application de règles (rapide)
Raisonnement	Non déductif	déductif
Acquisition	Facile (épisode de résolution d'un problème)	Difficile (comment faire pour résoudre un problème)

Table 1-1 : Comparaison entre RAPC et RAPR (Idiri et Napoli, 2012).

L'approche de raisonnement automatique par règles d'inférence présente un intérêt accru pour la problématique de surveillance maritime. Elle est utilisée pour l'identification des comportements à risques par l'implémentation du raisonnement humain. Cette approche est intéressante car elle est facile à mettre en œuvre si nous la comparons avec les modèles mathématiques par exemple. Un problème complexe peut être décomposé en un ensemble de règles. De plus, cette modularité sous forme de règles facilite la maintenabilité du système. En effet, les connaissances sont mises à jour facilement et rapidement dans un système de raisonnement alors qu'il est difficile par exemple de faire évoluer un modèle mathématique.

Pour bien comprendre les systèmes de raisonnement à base de règles, une présentation de son fonctionnement est exposée ci-après. Dans un système de raisonnement à base de règles, trois composantes essentielles sont définies : une base de connaissance ; les faits et un moteur d'inférence. Par analogie au raisonnement humain, la base de connaissance est l'ensemble de connaissances d'un être humain. Les faits sont la perception de son environnement (vue, goût, toucher, etc.). Le moteur d'inférence est le raisonnement humain. Comme nous le voyons sur la Figure 1-10, le moteur d'inférence vérifie en continu dans les événements en entrée (base de faits) s'il y a des règles applicables dans la base de règles. Avant d'exécuter ces règles, le moteur doit résoudre les conflits qui peuvent apparaître entre les règles applicables. Après l'exécution des règles sélectionnées par le moteur, de nouvelles règles et/ou de nouveaux faits vont venir enrichir respectivement, la base de règles et la base de faits (faits inférés, par exemple retourner une alerte).

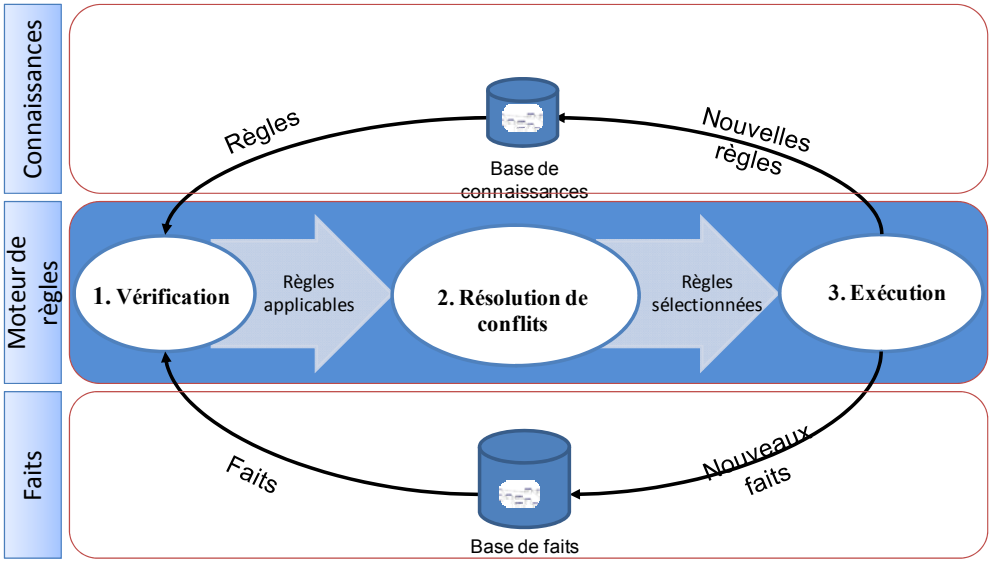


Figure 1-10 : Processus de raisonnement à base de règles
(Idiri et Napoli, 2012).

L'équipe de Jean Roy (Roy 2010) a utilisé le raisonnement à base de règles pour l'identification automatique de comportements anormaux liées à la sécurité maritime. Ils ont décrit un système complet de détection automatique de comportements anormaux, de la constitution de la base de connaissances à l'évaluation de menaces. Les connaissances expertes ont été définies sous forme de règles au cours d'un Workshop organisé au Canada avec des experts du domaine maritime. Ces règles décrivent des situations anormales de navires, connues auparavant par les experts qu'elles portent sur des informations statiques (signal AIS, numéro IMO, etc.) ou dynamique (vitesse, position, équipage, etc.). Quelques exemples de ces règles sont représentés sur la Figure 1-11. Une taxonomie des anomalies a été proposée sous forme d'ontologie (Roy, 2008) au cours de ce workshop. Studer (Studer et al., 1998) avait défini une ontologie comme étant « une spécification formelle et explicite d'une conceptualisation partagée». Pour plus de détail sur les ontologies, voir la section 1.6.2.

Après avoir conçu et développé le système de raisonnement à base de règles, il est possible d'identifier automatiquement les anomalies à partir des flux continus d'informations sur les navires. Concernant l'acquisition des connaissances expertes, J. Roy et son équipe se sont basés sur le brainstorming.

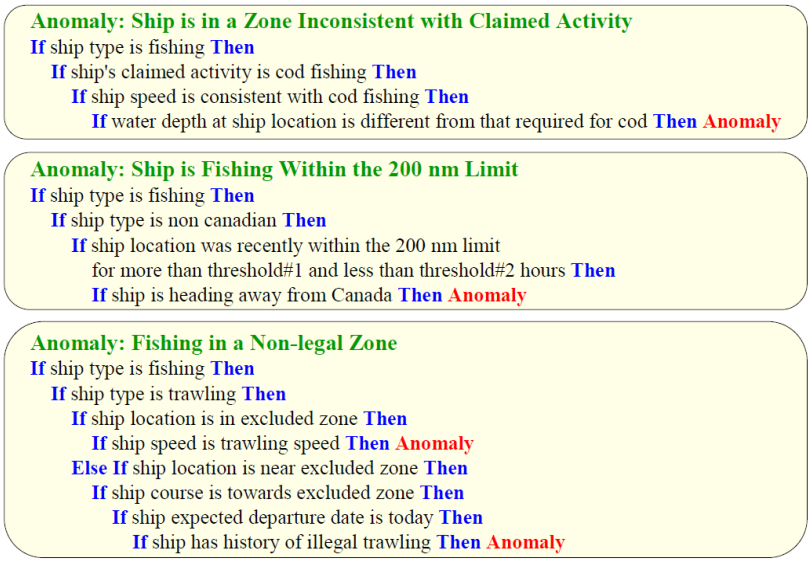


Figure 1-11 : Exemple de règles définies par brainstorming au Workshop organisé au Canada (J. Roy 2010).

Dans le projet SISMARIS, piloté par DCNS[35] et auquel a participé notre centre de recherche, un système de raisonnement à base de règles est aussi proposé pour analyser le trafic maritime sur des zones étendues. Les règles de connaissances intégrées au système sont issues des échanges avec les experts du domaine maritime comme les CROSS et la gendarmerie maritime.

Le point fort du raisonnement à base de règles est sa facilité de compréhension, sa modularité sous forme de règles (si condition alors action), sa maintenabilité et sa rapidité de traitement. Son point faible est lié à la difficulté d'acquisition et de capitalisation des connaissances (Estevez et al., 2006). L'acquisition des connaissances est considérée comme le goulot d'étranglement de cette méthode d'analyse de comportements.

1.6.2. Modélisation ontologique

L'analyse de comportements d'objets mobiles a une forte composante spatiale ce qui justifie l'intérêt des ontologies spatiales. Les ontologies spatiales présentent une potentialité intéressante pour l'analyse de comportements en intégrant le raisonnement à base de cas. Pour le raisonnement à base de cas, il a été défini dans la section 1.6.1. Une

[35] Groupe Français travaillant dans l'armement naval et l'énergie, http://fr.dcnsgroup.com/

ontologie est un ensemble de formalismes et de structures pour la formalisation et l'exploitation de connaissances d'un domaine. Les concepts et les termes partagés de ce domaine vont être organisés sous forme d'une hiérarchie de concepts qui est une structure de données en graphe décrivant des concepts partagés et leurs différentes relations. Les relations peuvent être d'inclusion comme le fait « chalutier » inclus dans le concept « navires de pêche » ou sémantiques comme la relation de voisinage ou de proximité.

Le but de cette modélisation ontologique est de formaliser le sens des termes et concepts d'un domaine pour pouvoir les exploiter par les personnes et les ordinateurs conjointement.

L'apport des ontologies pour l'analyse de comportements anormaux de navires a été évalué par Vandecasteele (Vandecasteele 2012) en développant un prototype appelé *OntoMap*. Ce prototype a permis de capitaliser des connaissances expertes et d'identifier automatiquement quelques comportements anormaux sur des données de déplacements de navires. Les informations maritimes utilisées dans ce prototype sont regroupées en trois ontologies : géométrique (position, trajectoires, etc.), cartographique et spécifique au domaine d'application (Vandecasteele 2012). L'ontologie métier, spécifique au domaine d'application a été constituée à l'aide de connaissances expertes issues de la littérature et d'interviews avec les experts (Vandecasteele 2012)[36].

Le système OntoMap (Figure 1-12) permet de définir des scénarios (cas résolus) en fonction d'un ensemble de propriétés et une distance sémantique à partir de laquelle le système décide de faire correspondre un ensemble de faits à un ou plusieurs scénarios. Prenons l'exemple de deux ou plusieurs navires qui naviguent à proximité dans une zone de pêche. Ces faits retournent des alertes associées aux navires concernés. Si la distance sémantique entre les faits et un scénario prédéfini est petite alors il y a de fortes chances que le scénario corresponde.

[36] Page 39

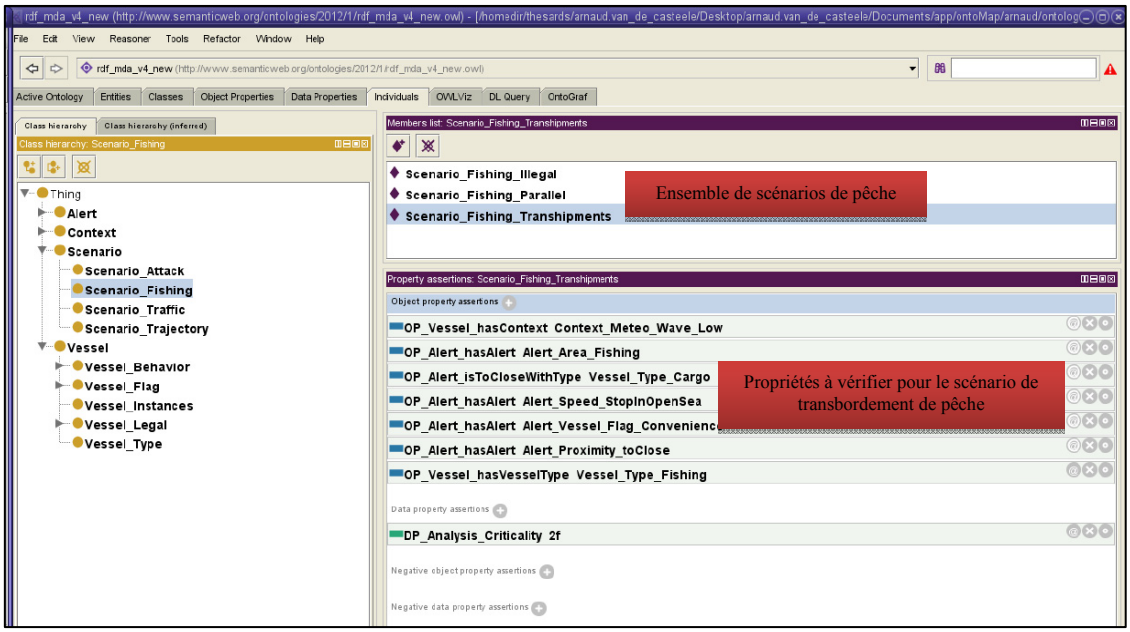


Figure 1-12 : Capture d'écran du système OntoMap (Vandecasteele 2012).

1.6.3. Modélisation par Classifieur Bayésien

La classification Bayésienne naïve est une méthode d'apprentissage supervisée qui met en œuvre des classifieurs pour la reconnaissance de formes, la prédiction et le tri. Cette méthode est basée sur le théorème de Bayes et permet de mettre en évidence des combinaisons linéaires entre les observations. L'appellation « naïve » est attribuée à cause de l'hypothèse de départ concernant l'indépendance des caractéristiques des objets classés. Prenons l'exemple d'une embarcation considérée comme étant de classe zodiac si la longueur et la vitesse mesurées appartiennent à des classes de valeurs prédéfinies. Ces deux descripteurs, à savoir la longueur et la vitesse sont supposés être indépendants. Malgré cette hypothèse simpliste de départ, ce type de classifieur est très utilisé et se révèle efficace et robuste.

En se basant sur cette méthode de classification, le projet Système d'Alerte et de Réponse Graduée Offshore (SARGOS), financé par l'agence ANR[37], propose comme son nom l'indique, une réponse à l'encontre d'actes de malveillance comme la piraterie, le terrorisme auxquels les infrastructures offshore sont vulnérables. Il supporte toute la chaîne de traitement, de l'identification de la menace à la proposition d'une réponse

[37] Agence Nationale de Recherche, structure gouvernementale Française qui finance les recherches publiques

graduée. A partir d'informations récupérées par des radars à ondes continues (FMCW[38]) installés sur des plateformes offshores, les petites embarcations vont être détectées et classifiées en se basant sur les classifieurs Bayésiens. Dans le cadre de l'expérimentation de cette méthodologie, l'apprentissage s'est fait sur des données d'observation collectées à partir d'un radar FMCW installé sur le site de la Direction Générale de l'Armement (DGA) de Saint-Mandrier (Giraud et al., 2013). L'apprentissage permet de construire un dictionnaire composé de matrices de vecteurs forme regroupant les caractéristiques de chaque classe de navire (TéSA 2011). Chaque vecteur forme décrit la forme, la géométrie et la topologie d'une observation cible. C'est l'utilisation en situation opérationnelle de ce dictionnaire qui va permettre la classification automatique de nouvelles observations. Dès qu'un objet pénètre dans le périmètre du radar, il est identifié et classé (Figure 1-13).

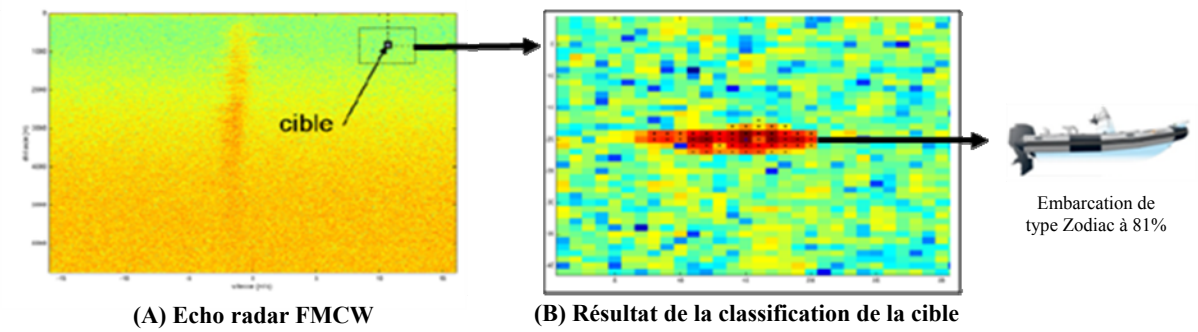


Figure 1-13 : Ciblage et classification d'un comportement d'intrusion d'une embarcation à partir d'images radar Range-Doppler d'un FMCW (TéSA 2011).

Pour chaque nouvelle observation détectée, un vecteur forme est mesuré puis comparé aux vecteurs formes se trouvant dans le dictionnaire pour prédire son type. Une distance sémantique est alors calculée entre ce vecteur forme et les vecteurs forme du dictionnaire pour trouver la classe d'appartenance la plus probable. Les descripteurs de l'objet choisis dans le vecteur forme doivent être discriminant entre les objets appartenant à des classes différentes et confondant entre les objets de même classe. Il peut cependant s'avérer difficile de classifier automatiquement des objets ayant des formes qui peuvent se confondre.

[38] Frequency Modulated Continuous Wave

1.6.4. Autres méthodologies

On trouve aussi dans la littérature, d'autres travaux de recherche sur l'analyse de situations à risques de navires par des approches probabilistes (Amrozowicz et al., 1997) et fondées sur la simulation numérique (Fournier 2005) (Nishizaki et al., 2011). Ces approches ne considèrent pas les historiques de données qui sont riches en enseignements, ce qui rend la découverte de nouvelles connaissances difficile.

1.7. Limites des méthodologies de modélisation actuelles

La plupart des méthodologies de modélisation des risques maritimes proposées actuellement sont basées sur la formalisation de connaissances expertes, qu'elles soient issues de revue de littérature, du brainstorming ou d'interviews avec les experts. Les connaissances issues de ces méthodes dépendent de l'expérience des experts et sont donc subjectives du fait qu'elles peuvent changer d'un expert à l'autre. De plus, ces méthodes ne permettent pas la découverte de connaissances nouvelles décrivant des comportements auxquels les experts n'ont jamais été confrontés.

Selon N. Sumpter (Sumpter & J. Bulpitt 2000), la modélisation de comportements d'objets par apprentissage sur des observations peut produire de meilleurs modèles de comportement, plus réaliste que la modélisation à base d'expertise. L'apprentissage automatique ou la fouille de données permet de bien capturer les caractéristiques réelles à partir des observations et la découverte de nouvelles connaissances. Elle utilise des outils issus de plusieurs domaines comme l'intelligence artificielle, la théorie de l'apprentissage, la théorie de l'information et les statistiques qu'elles soient inférentielles ou descriptives. Elle intègre donc des outils élaborés permettant la découverte de connaissances complexes, ne pouvant être découvertes en tâtonnant par exemple sur des résultats de statistiques descriptives (Tufféry 2010).

1.8. Synthèse sur les méthodes d'analyse et de modélisation de comportements de navires

Dans les sections précédentes, des méthodologies d'analyse et de modélisation de comportements de navires ont été décrites. Dans l'objectif de synthétiser ces méthodologies, deux tableaux récapitulatifs sont présentés ci-dessous.

	Description	**Avantages**	**Limites**
Analyse statistique	Intègre deux approches (exploratoire et confirmatoire) pour la description synthétique des variables, la recherche de dépendances, la mesure de déviations, la découverte de relations de causalité, etc.	- Etudier les distributions des variables univariées et multivariées, - Infirmer ou confirmer une hypothèse, - Inférence des conclusions à la population entière avec une marge d'erreur calculée, - Utiliser des représentations graphiques intuitives (Box Plot, histogramme, etc.).	- Nécessite de poser des hypothèses de départ, - Ne permet d'analyser qu'une dizaine de milliers d'individus et quelques dizaines de variables.
Analyse visuelle	Utilise plusieurs formes de visualisation de données pour permettre l'analyse et l'extraction visuelle de connaissances.	- Améliorer les capacités de perception, compréhension et raisonnement des utilisateurs, - Intégrer la dimension humaine dans l'exploration de données pour faire participer les utilisateurs dans la construction des connaissances.	- Raisonnement humain sur les différentes formes visuelles requiert un temps prohibitif pour la construction de connaissances, - Complexité de certaines visualisations et leur inadaptation par rapport aux données analysées, le type d'analyse et le profil de l'utilisateur rend difficile l'interprétation des résultats.
Fouille de données	Combine plusieurs techniques statistiques, d'intelligence artificielle, d'algorithmiques, d'informatiques, etc. pour l'extraction automatique de connaissances à partir de bases de données.	- Générer automatiquement des hypothèses (meilleure productivité), - Découvrir de nouvelles connaissances, - Très grandes capacités d'analyse en termes de nombre d'individus et de variables, - Découvrir des modèles plus réalistes que les modèles à base d'expertise,	- Ne fait pas participer les utilisateurs dans l'exploration des données, - Les résultats demandent de la compétence en fouille de données pour les interpréter et les valider.

Table 1-2 : Synthèse des méthodes d'analyse utilisées dans l'étude de comportements de navires

	Description	Avantages	Limites
Règles d'inférence	Formalisation de connaissances sous forme de règles de conjonction de disjonction d'occurrences : « ***si conditions alors actions »***	- Facilité de compréhension par les utilisateurs, - Modularité des règles, - Réalisation de raisonnements déductifs, - Facilité de mise en œuvre du système modélisé et sa maintenabilité dans le temps.	- Nombre important de règles de connaissances à gérer, - Difficulté d'acquisition de la base de connaissances (comment faire pour résoudre un problème ?).
Modélisation Ontologique	Formalisation de connaissances sous forme de structures en hiérarchies de concepts et de formalismes pour les partager et les exploiter par les humains et les machines conjointement.	- Partage de connaissances entre les utilisateurs ; entre les utilisateurs et les systèmes et entre les systèmes, - Réalisation de raisonnement automatique sur les ontologies, - Facile à appréhender par les utilisateurs à cause de sa représentation en réseau de graphes.	- Dépendance forte avec le problème à résoudre, - Insuffisance des formalismes exploitant la dimension spatiale.
Classifieurs Bayésiens	Modèles mettant en exergue des combinaisons linéaires entre les variables pour la reconnaissance de formes, la prédiction et le tri par exemple.	- Efficacité des classifieurs Bayésiens, - Graduation des résultats (modèle probabiliste).	- Condition de départ sur l'indépendance des variables caractérisant les objets classés, - Difficulté de classer automatiquement des objets qui se confondent.

Table 1-3 : Synthèse des méthodes de modélisation utilisées pour la formalisation de comportements de navires

1.9. Conclusion

L'analyse de mouvements d'objets mobiles, leur compréhension, la compréhension des interactions entre les objets et entre les objets et l'environnement présentent des perspectives intéressantes pour des domaines d'application très variés comme la surveillance du trafic maritime.

Identifier des comportements inhabituels, à risques, appréhender les objectifs et les contraintes d'évolution à partir des mouvements est une avancée importante pour la surveillance automatique.

La plupart des travaux de modélisation pour l'identification automatisée des comportements inhabituels, anormaux ou suspects de navires proposés dans la littérature sont certes intéressants mais présentent des limites. Dans ces travaux, la méthode de brainstorming est souvent utilisée. Les scénarios produits dépendent alors beaucoup de l'expérience personnelle de chaque expert. De plus, elle ne permet pas la découverte de connaissances nouvelles. Dans la littérature, la modélisation des risques maritimes par fouille de données automatique est peu explorée (Darpa 2005) alors qu'elle peut combler les limites actuelles.

La fouille de données permet d'extraire des caractéristiques de comportements cachées et réelles à partir d'observations passées ce qui peut aboutir à de meilleurs modèles de comportements.

Dans le chapitre suivant, nous présentons un état de l'art des domaines de fouille de données et nous identifions les méthodes permettant d'extraire des connaissances sur les comportements à risques de navires.

Chapitre 2 : Contribution de la fouille de données à l'analyse de comportements

2.1. Introduction

La communauté des chercheurs en fouille de données (*data mining)* est très active. Plusieurs travaux de fouille de données, fouille de données spatiales et/ou temporelles (Roddick et al., 2001) ont vu le jour dans des domaines d'application très variés comme la gestion de la relation client (identification de prospects, churn[39], etc.), le marketing stratégique (mailing[40], association de produits, etc.), la gestion des risques (remboursement de crédits, détection de fraudes, etc.) et la prévision du trafic routier. L'acquisition de connaissances par induction a donc démontré son succès par son large panel d'applications. La fouille de données est passée rapidement du cadre de la recherche vers l'industrie pour occuper une place importante dans le domaine de l'aide à la décision (Tufféry 2010)[41].

La fouille de données quel que soit son domaine, classique, spatial, d'objets mobiles ou du trafic peut contribuer à l'analyse des comportements d'objets en déplacements. Les situations et les mouvements intéressants de mobiles peuvent être mis en exergue en détectant des relations et des motifs cachés décrivant des comportements.

Ce chapitre est organisé en trois parties : la première partie décrit les domaines de la fouille de données à savoir la fouille de données classique, la fouille de données spatiales, la fouille de données d'objets mobiles et la fouille de données de trafic d'objets. La deuxième partie, expose deux prototypes d'analyse de comportements basés sur la fouille de données. Enfin la troisième partie, présente une synthèse des méthodes et algorithmes de fouilles de données.

2.2. Les domaines de la fouille de données

Les volumes importants de données générés actuellement, offrent un potentiel important à la fouille de données. Les bases de données peuvent contenir de nombreuses informations d'intérêt, cachées par le volume important de données. Il est naturel de vouloir exploiter ces données et extraire ces informations. Cependant, le volume de données et leur exploitation n'évoluent pas de la même manière. Les données enregistrent une croissance exponentielle alors que leur exploitation est linéaire, ce qui engendre des

[39] Identification de clients qui sont susceptible de partir à la concurrence.
[40] Le publipostage est l'adaptation de la communication aux différents segments de clientèle.
[41] Page 13.

pertes d'informations non-négligeables. Avec les méthodes d'analyse statistiques, il est difficile de traiter des millions d'enregistrements, avec plusieurs centaines de variables ayant des types de plus en plus complexes (texte, géométrie, réseau, etc.). Ces méthodes atteignent leurs limites face aux volumes de données de plus en plus importants. Pour faire face à ce problème, la fouille de données a été inventée. Les enjeux actuels liés à la concurrence du marché, le besoin de rapidité de traitement des données, de prise de décision ont poussé les industries à s'approprier la fouille de données rapidement (Tufféry 2007)[42].

Quant à ce qui l'a fait émerger, il s'agit du développement des moyens de stockage et de calcul informatique, l'évolution du domaine décisionnel et la possibilité de prendre en compte les bruits présents dans les données collectées (données lacunaires). La fouille de données a rapidement été utilisée dans plusieurs domaines qui partent de l'infiniment petit comme la génomique à l'infiniment grand comme l'astrophysique (Tufféry 2007)[43]. Etant donné les potentialités offertes par la fouille de données, son spectre d'application est très large. Elle est utilisée en commerce pour l'analyse des comportements de consommation par exemple, dans les banques pour la distinction entre les bons et mauvais payeurs et en assurance pour l'identification des fraudes et les critères explicatifs des risques.

Selon la nature des objets étudiés, plusieurs domaines de fouille de données peuvent intervenir. Nous présentons ci-après un panorama des méthodes de fouille de données organisées par domaine, à savoir la fouille de données classique, spatiales, d'objets mobiles et du trafic. La fouille de données classique s'intéresse à l'exploration de données relationnelles, la fouille de données spatiales s'intéresse à l'exploration de données spatiales et la fouille de données d'objets mobiles et de trafic de mobiles sont des domaines plus récents qui s'intéressent à l'exploration de données de capteurs de mobiles.

[42] Page 13
[43] Page 1

Aujourd'hui, énormément de problèmes, de méthodes et d'algorithmes de fouille de données existent dans la littérature. Chaque domaine possède une panoplie de problèmes d'extraction de connaissances, chaque problème détient plusieurs méthodes et chaque méthode dispose de plusieurs algorithmes comme illustré à la figure 2-1.

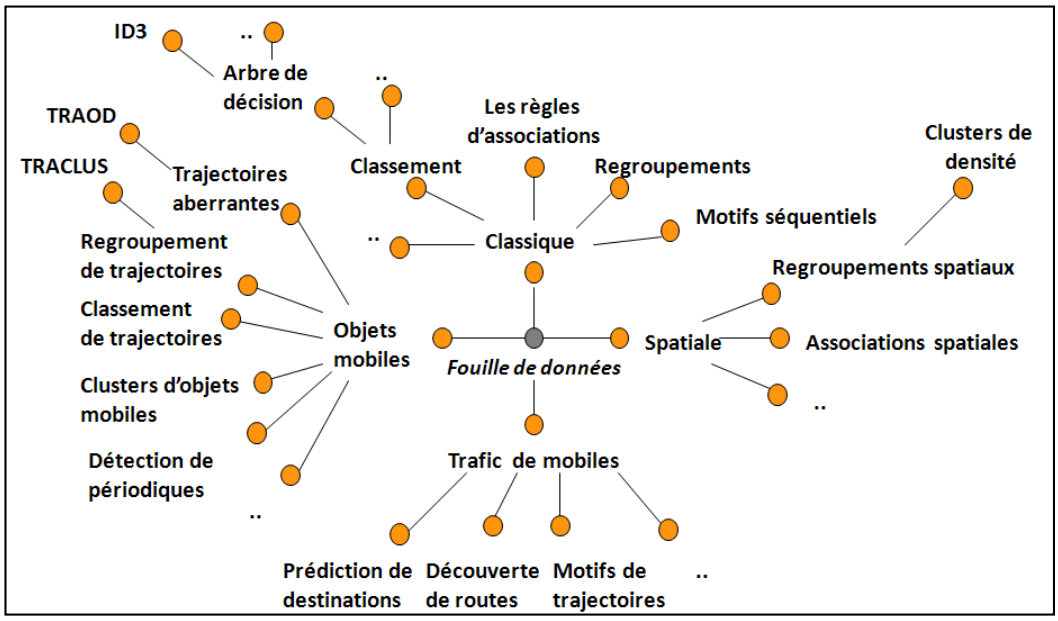


Figure 2-1 : Organisation de l'état de l'art de la fouille de données.

Nous allons présenter dans les sous sections suivantes, un état de l'art sur la fouille de données. Dans cet état de l'art, nous détaillerons par domaine les méthodes et algorithmes nécessaires pour la compréhension de ce travail.

2.2.1. La fouille de données classique

Nous appelons fouille de données classique, l'ensemble des problèmes de fouille de données qui ne considèrent pas les relations spatiales. Les objets étudiés sont souvent des tuples de bases de données ne contenant pas de dimension spatiale (localisation absolue). Les tuples peuvent être définis comme des ensembles de valeurs d'attributs ordonnées relatifs à un enregistrement (observation). Ces tuples sont organisés sous forme de structures de données en table[44] qui sont reliées par des relations logiques. Pour ce qui est des relations spatiales, elles vont être définies dans la section 2.2.2.

[44] Une table en base de données est une structure en tableau où les lignes sont des enregistrements et les colonnes sont des attributs

Nous allons exposer dans les sous-sections suivantes, les différents problèmes de la fouille de données classique : les associations, le classement, le groupement, la fouille des séries chronologiques et l'analyse des aberrations.

2.2.1.1. Les associations

Les associations sont des problèmes non supervisés de fouille de données permettant d'extraire des relations d'implication au sein d'un même événement ou entre des séquences d'événements ordonnés dans le temps. Nous allons présenter ci-après, les deux problèmes d'association qui existent, à savoir, la recherche de règles d'association et la recherche de motifs séquentiels.

2.2.1.1.1. Les Règles d'association

Les règles d'association sont des règles d'occurrence mettant en exergue les relations cachées par le volume important des bases de données. Elles sont basées sur la découverte de motifs fréquents qui sont des ensembles d'items apparaissant souvent ensemble dans les bases de données. Un item étant une combinaison entre un attribut et une valeur. Formellement, une base de données D est définie par un triplet (O, P, R) tel que O est l'ensemble fini des objets de la base (appelé individus), P l'ensemble fini d'attributs (appelés variables) et R une relation binaire entre les deux ensembles O et P. Etant p un itemset appartenant à P, on dit que o contient p si (o,p) appartient à R, tel que o appartient à O. La découverte de règles d'association consiste à sélectionner tous les ensembles d'items I inclus dans P qui sont fréquents dans O. Ces ensembles d'items appelés itemsets sont utilisés pour générer les règles d'association r de la forme $\boldsymbol{A \rightarrow B}$ tel que : ***A*** et ***B*** sont deux itemsets fréquents et leur intersection est vide. ***A*** est appelé antécédent de la règle ou partie gauche et ***B*** est appelé conséquent ou partie droite. Dans un processus d'extraction de règles d'association, seuls les itemsets ayant un bon support et les règles ayant une bonne confiance sont conservées. Le support est un indicateur de fiabilité de la règle, appelé aussi porté. Il est égal à la fréquence d'apparition d'un itemset dans les transactions par rapport au nombre total de transactions de la base de données (Tufféry 2007). La mesure support est évaluée comme suit :

Supp(A → B)= card(AB)/card (O), *card* est le cardinal de l'ensemble ou la fréquence d'apparition.

La confiance représente l'indicateur de précision de la règle qui est égal à la fréquence d'un itemset par rapport à la fréquence d'apparition de la première partie de la règle d'implication (Tufféry 2007). La mesure confiance est évaluée comme suit :

Conf(A → B)= card(AB)/card (A).

Le principe est le suivant, les supports et les confiances devraient être calculés pour toutes les règles possibles puis comparés aux seuils "minsupp" et "minconf" qui sont définis *a priori* par les utilisateurs. Les règles ayant les valeurs des paramètres support et confiance au-delà des seuils donnés par les utilisateurs sont conservées car elles possèdent des relations dites fortes. Le coût important de la recherche de toutes les règles a poussé à découpler les conditions de support et de confiance[45]. Le problème est traité le plus souvent en deux phases :

1. Génération d'itemsets satisfaisant les conditions de support minimal,
2. Génération des règles satisfaisant les conditions de confiance à partir d'itemsets précédemment générés.

Il existe d'autres mesures proposées dans la littérature pour améliorer la sélectivité car trop de règles sont extraites dont certaines sans intérêt. Parmi ces mesures, nous pouvons citer les mesures Lift, Pearl, Loevinger, Surprise et J-mesure. Nous ne pouvons pas exposer toutes les mesures dans ce manuscrit, mais nous renvoyons pour plus d'informations, aux travaux de S. Lallich (Vaillant et al., 2005) (Le Bras et al., 2011) qui traite en détail cette question.

Une mesure de performance souvent utilisée avec le support et la confiance pour évaluer l'amélioration apportée par une règle par rapport au hasard est la mesure Lift. Le résultat de cette mesure permet d'identifier la corrélation entre les deux parties de la règle d'association :

- Un Lift supérieur à 1 indique une corrélation positive,
- Un Lift égal à 1 indique une corrélation nulle,
- Un Lift inférieur à 1 indique une corrélation négative.

La mesure Lift est évaluée comme suite :

[45] Valérie Monbet, Université Rennes 1,
Lien http://perso.univ-rennes1.fr/valerie.monbet/doc/cours/IntroDM/Chapitre5.pdf

Lift(A → B)= Conf(A → B)/card (B).

L'exemple du panier de la ménagère (Agrawal et al., 1993) illustre bien la recherche de règles d'association. Une base de données de transactions (les paniers) est composée d'items (les produits achetés). La découverte des associations consiste à chercher les itemsets (ensemble d'items) fréquemment liés dans une même transaction. Les règles extraites sont par exemple « {bière}→{couches} (30%, 80%) : 80% des gens qui achètent de la bière, achètent également des couches et ces clients représentent 30% des cas». Cette règle est entre items singletons et les règles entre plusieurs items peuvent être plus susceptibles d'impressionner car elles sont plus difficiles à identifier. En sachant quels items sont fréquemment achetés ensemble, une stratégie marketing peut être développée. La découverte de relations entre les items permet par exemple d'organiser les rayons en mettant les produits fréquemment achetés ensemble à proximité pour augmenter les ventes.

Il y a toute une batterie d'algorithmes d'extraction de règles d'association qui existent comme Apriori, FP-growth, TreeProjection. Ces algorithmes différents du point de vue des performances (temps et espace requis d'exécution). FP-growth (Han et al., 2000) par exemple améliore la méthode Apriori en utilisant une représentation des itemsets sous forme d'index appelé FP-Tree (Frequent Pattern Tree) pour éviter de scanner la base de données plusieurs fois. Gardarin (Gardarin et al., 1998) propose aussi une amélioration basée sur un index bitmap pour accélérer le calcul des supports.

Apriori (Finding Frequent Itemsets Using Candidate Generation) (Agrawal & Srikant 1994) est l'algorithme fondateur de l'extraction de règles d'association. Cet algorithme commence par une phase de recherche des itemsets fréquents (candidats) en balayant la base de données. Une structure en treillis permet de faire une génération des itemsets par niveau (itemsets de longueurs *1, 2,..,k*, *k* étant la longueur maximale des itemsets de la base de données). Un itemset est considéré comme fréquent si le support calculé est supérieur ou égal au support minimal fixé par l'utilisateur (condition de support). Dans la deuxième phase, Apriori sauvegarde les règles de type $A \rightarrow B$ dont la confiance dépasse la confiance minimale fixée par l'utilisateur (condition de confiance). A et B sont des itemsets fréquent et $A \cap B = \{\}$. Pour plus de détail, voir l'algorithme *Apriori* (Algorithme 2-1).

Algorithme Apriori

Entrée : minsupp, minconf, D (base de données)
Sortie : L (ensembles d'items fréquents), R (règles d'association)

Début
Phase I :
K=1 ; L=∅ ; C1= {Candidats de taille 1} ; //initialisation
L1=Frequent(1,C1) ; L=L∪L1 ; //garder les items fréquents
Tant que Lk=∅
K++ ;
Ck = Candidats(K,Lk-1); //générer les itemsets de taille K
Lk = Frequent(K,Ck) ; L=L∪Lk ;//garder les items fréquents
Fin Tant que;

Phase II :
R=∅;
Pour chaque ensemble I de L
Pour chaque sous-ensemble S non vide de I
Si Conf (S → I-S) >= minconf
r= " S → (I-S) " ;
R=R ∪ {r} ;
Fin Si

Fin

Algorithme 2-1 : Algorithme Apriori.

Pour des raisons d'optimisation et de réduction de l'espace de recherche, les données sont organisées dans un treillis de Galois. En effet, cette structure de données permet d'ordonner l'ensemble des itemsets selon la relation d'inclusion ensembliste et d'exploiter la fermeture du support (Antimonotonie). L'antimonotonie de la condition de support est une propriété qui va aider à l'élagage du treillis en supprimant les branches qui n'ont pas le support, c'est à dire les itemsets qui ont un support s% inférieur au seuil "minsupp" défini *a priori* par les utilisateurs (voir section 2.2.1.1.1). En effet, si le support d'un ensemble est de s% alors le support de tout sous-ensemble de cet ensemble est au moins égal à s%. La contre-apposée donne, si un ensemble n'a pas le support alors tous les ensembles supérieurs contenant cet ensemble ne l'auront pas aussi. Ce résultat va réduire massivement l'ensemble des itemsets en élaguant toutes les branches supérieures du treillis qui sont reliées aux ensembles n'ayant pas le support.

Nous présentons sur la Figure 2-2 un treillis de Gallois ordonné par niveau où chaque niveau k compte tous les ensembles formés par la combinaison de k itemsets (itemset de longueur k). Il est aisément remarquable que le nombre d'itemsets dans le treillis augmente d'une manière combinatoire.

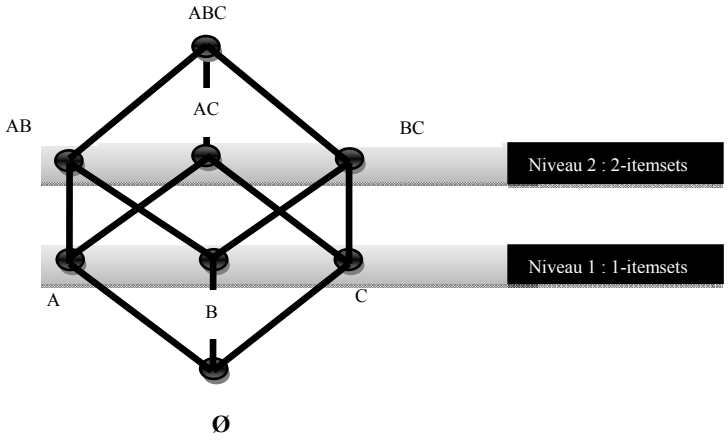


Figure 2-2 : Treillis de Galois.

2.2.1.1.2. Les Motifs séquentiels

Le problème de fouille de motifs séquentiels a été introduit pour la première fois par Srikant et Agrawal (Srikant & Agrawal 1996). La recherche de motifs séquentiels peut être vue comme une extension de la notion de règles d'association, intégrant des contraintes temporelles. Cette recherche met en évidence des associations entre les transactions alors que la recherche des règles d'association détermine les liens au sein d'une même transaction. Les motifs séquentiels sont extraits à partir de séquences d'événements ordonnées et souvent sauvegardés dans des bases de données transactionnelles (voir l'exemple de la Table 2-1). Ces événements ne contiennent pas forcement la notion de temps d'une manière concrète mais elle est implicite dans l'ordre des événements (Han & Kamber 2006)[46]. Une règle séquentielle dont la notion de temps est implicite peut être « 60% des navires qui prennent un tronçon de chemin A, arrivent au point B ». Si la temporalité est présente d'une manière explicite dans les transactions et l'algorithme utilisé intègre les contraintes temporelles, la règle peut devenir «60% des navires qui prennent un tronçon de chemin A, arrivent au point B **dans les 2 heures qui suivent** ».

[46] Page 498

Identifiant_séquence	Séquence
1	<*a(**ab**c)(a**c**)d(cf)*>
2	<*(ad)c(bc)(ae)*>
3	<*(ef)(**ab**)(df)**c**b*>
4	<*eg(af)cbc*>

Table 2-1 : Exemple de base de données séquentielles. La base est au format horizontal, elle contient 4 séquences d'événements, la 1ère séquence contient 5 évènements et le 2ème événement de cette séquence a un itemset (ab)c de taille 3. La recherche de la fréquence d'apparition de cette séquence dans la base donne 2 (itemset en rouge gras) ((Han & Kamber 2006), page 499).

Dans certains contextes, l'identification des événements d'individus au cours du temps est indispensable afin de pouvoir suivre leurs comportements séquentiels (Srikant & Agrawal 1996). Plusieurs applications utilisent ce type de problème pour l'extraction des comportements séquentiels comme l'analyse de séquences biologiques, l'analyse des séquences naturelles et l'analyse des comportements des utilisateurs qui naviguent sur le web (enchaînement des clics de souris sur une page web, etc.)

Plusieurs algorithmes d'extraction de motifs séquentiels sont proposés dans la littérature, nous pouvons citer SPADE (Zaki 2001), GPS (Srikant & Agrawal 1996) et PrefixSpan (Pei et al., 2001). Comme Apriori, la majorité des algorithmes d'extraction de motifs séquentiels se basent sur la fouille des itemsets fréquents vu dans la section 2.2.1.1. Chacun des algorithmes cités précédemment, représente une approche d'extraction particulière. SPADE par exemple, fait de la génération de candidats à partir d'un format vertical de la base de transactions où chaque itemset contient les identifiants de la séquence et de l'événement des occurrences des itemsets présents dans la base *(<itemset : (identifiant_séquence, identifiant_événement)>*). GPS suit presque la même approche mais l'extraction se fait à partir d'un format horizontal (*<Identifiant_séquence : séquence_itemsets>*) (comme le format de Table 2-1). Enfin, PrefixSpan adopte plutôt une approche de comptage de chemins d'une structure de nœuds d'items (pattern growth method) qui se rapproche de la méthode FP-growth pour la découverte de règles d'association (Cf. section 2.2.1.1.1).

Masseglia a fait un état de l'art intéressant sur les méthodes d'extraction de motifs séquentiel (Masseglia et al., 2004).

2.2.1.2. Le classement et prédiction

Selon S. Tufféry (Tufféry 2007), « le classement estime la valeur d'une variable *à expliquer* par d'autres variables du même individu appelées *cibles* ou *explicatives*. Si la variable à expliquer est qualitative alors la technique est appelée Classement et si elle est continue, elle est appelée Prédiction ». Le classement consiste à analyser de nouvelles données et à les affecter à des classes prédéfinies ou modélisées au préalable. Le classement et prédiction sont des problèmes supervisés d'analyse de données qui sont souvent utilisés pour prédire des valeurs ou des libellés de classes.

Plusieurs techniques, approches et méthodes de classement et de prédiction existent, comme les arbres de décision, les réseaux bayésiens, le raisonnement à base de règles, les algorithmes génétiques, la régression linéaire/non linéaire et le support vecteur machine. Elles utilisent toutes deux étapes : une étape d'apprentissage et une étape de classement. La première étape se focalise sur la construction du classifieur qui va caractériser chaque classe importante par rapport aux données d'apprentissage (variables cibles) et l'attribut des libellés de classes (variable à expliquer). La deuxième étape consiste quant à elle à classifier les nouvelles données selon le classifieur défini dans la première étape.

Bien qu'elles utilisent les mêmes phases, les méthodes peuvent donner des résultats différents qui peuvent être liés aux caractéristiques des données analysées. La comparaison entre les différentes méthodes est parfois donc nécessaire. La comparaison peut être sur la capacité qu'a la méthode à bien classer les nouvelles données (précision), la vitesse de génération et d'utilisation du classifieur, la capacité à donner de bons résultats malgré la présence de données bruitées (aberrantes, manquantes, etc.) et la scalabilité[47].

Les arbres de décision sont parmi les méthodes les plus utilisées à cause des règles explicites fournis par le classement, sa faible indépendance avec l'échantillon de données (robustesse), sa facilité de compréhension et d'interprétation par les utilisateurs. Cette méthode présente les résultats du classement sous forme d'une structure en organigramme. Chaque nœud de l'arbre est un attribut et chaque branche reliant les nœuds est une valeur de l'attribut ou un intervalle de valeurs si l'attribut est continu

[47] Capacité d'un système de monter en charge.

(Figure 2-3). L'intuitivité de cette méthode a fait son succès, plusieurs algorithmes implémentant cette méthode ont été développés. Certains de ces algorithmes ne produisent que des arbres binaires dont chaque nœud de l'arbre produit deux branches. Parmi les algorithmes les plus connus et utilisés, on trouve CART (Breiman et al., 1984), ID3[48] (Quinlan 1979; Quinlan 1982; Quinlan 1986) et C4.5 qui est le successeur de ID3. Contrairement à ID3, C4.5 prend en compte les attributs continus et les valeurs manquantes.

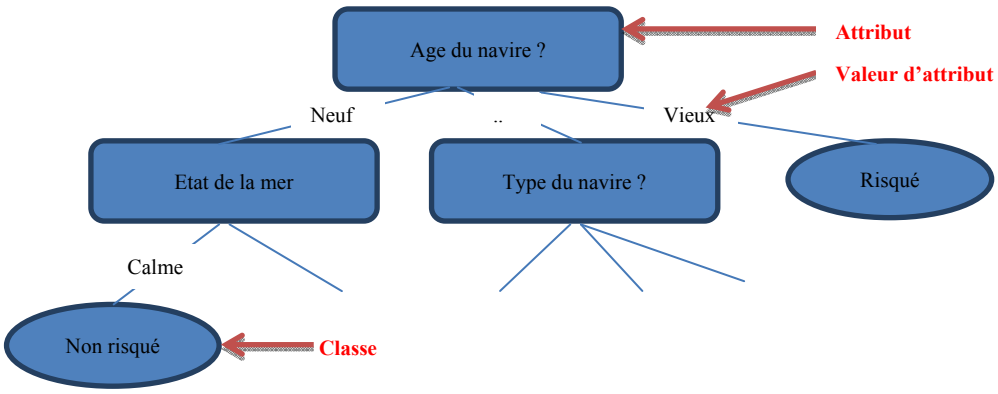


Figure 2-3 : Exemple d'arbre de décision

La majorité des algorithmes de classement par arbre de décision prennent en entrée, les données d'apprentissage avec la distinction de l'attribut de libellé de classe et la méthode de sélection d'attributs les plus discriminant. Cette méthode heuristique de sélection permet de déterminer les critères de séparation entre les classes en utilisant des mesures de sélection comme l'index de Gini et le gain d'information. L'objectif est de trouver les attributs qui séparent le mieux les classes pour avoir un modèle en arbre le plus concis possible.

Prenons l'exemple de l'algorithme ID3 (Algorithme 2-2), qui suit le principe suivant : il prend en entrée un échantillon de données E, un ensemble d'attributs A, les libellés de classes c et donne en sortie une racine d'un arbre de décision. L'algorithme calcule récursivement sur les attributs A non encore choisis par l'algorithme, l'entropie de Shannon pour trouver quel attribut maximise le gain d'information. Quand cet attribut est identifié, un nœud est créé. Si ce nœud est terminal il est étiqueté à Classe, sinon un autre attribut est sélectionné, un sous arbre est créé et l'algorithme passe à un nœud

[48] Interactive Dichotomiser 3

suivant qui n'a pas encore été exploré. Quand l'un des critères d'arrêt est vérifié, il s'arrête et retourne le nœud racine qui est le premier nœud de l'arbre.

L'entropie de Shannon d'un attribut S ayant i valeurs d'attribut $S(x_1,...,x_i)$, sachant que le nombre de classes est égal à j, $C(c_1,...,c_j)$ est calculée comme suit :

$E(S) = -\sum_i P(x_i) \sum_j P(c_j/x_i) \log (P(c_j/x_i))$, $P(c_j/x_i)$ est la fréquence relative de la classe j dans le segment S.

Algorithme ID3

Entrée : E (échantillon de données), A (ensemble d'attributs),
C (classes)
Sortie : Racine (Racine de l'arbre de décision)

Début
initialiser l'arbre à vide;
Si tous les exemples de E ont la même classe c
Alors étiqueter la racine par c;
Sinon si l'ensemble des attributs A est vide
Alors étiqueter la racine par la classe majoritaire dans E;
Sinon soit a le meilleur attribut choisi dans A;// *celui qui maximise le gain*
étiqueter la racine par a;
Pour toute valeur v de a
Construire une branche étiquetée par v;
Soit Eav l'ensemble des exemples tels que e(a) = v;
Ajouter l'arbre construit par ID3(A-{a}, Eav, c);
Finpour
Finsinon
Finsinon

Retourner racine;
Fin

Algorithme 2-2 : Algorithme ID3 proposé par Quainlan (Quinlan 1986).

2.2.1.3. Le groupement

Le groupement est utilisé depuis toujours dans le subconscient humain pour distinguer les différents éléments qui composent le monde qui l'entoure (Han & Kamber 2006)[49]. Un schéma conceptuel est créé et amélioré continuellement par apprentissage, pour distinguer par exemple un navire de pêche et un navire de plaisance et un risque d'un non-risque. Appelé Clustering en anglais, l'objectif de ce problème est de savoir grouper, dans les mêmes classes, les enregistrements (individus) qui semblent similaires. Contrairement au classement, le groupement n'a pas de variables à expliquer, c'est un

[49] Page 384

problème non-supervisé de fouille de données. En effet, le groupement n'a pas besoin de connaître à l'avance les classes auxquelles appartiennent les individus mais il peut les définir. Le classement par contre a besoin d'une phase de groupement si les classes des individus ne sont pas connues à l'avance.

Plusieurs techniques peuvent être utilisées pour la découverte de clusters. Ces techniques sont de plusieurs catégories : les méthodes de partitionnement, les méthodes hiérarchiques, basées sur la densité, basées sur les grilles, sur les modèles, etc. De nombreux algorithmes efficaces ont été proposés dans la littérature optimisant les performances et la qualité des classes obtenues dans de grandes bases de données (Tufféry 2007). Voici les plus couramment utilisés, organisés par méthodes :

- méthodes par partitionnement (K-means, K-medoids, CLARANS, EM, etc.),
- méthodes hiérarchiques (DIANA, BIRCH, CURE, etc.),
- méthodes de densité (DBSCAN, OPTICS, etc.),
- méthodes de grilles (STING, WaveClusters, Clique),
- méthodes par modèle (Réseaux de neurones, etc.).

Les méthodes par densité sont particulièrement intéressantes car elles permettent de gérer le bruit (données aberrantes, etc.), de découvrir des clusters ayant des formes arbitraires et d'identifier automatiquement le nombre de clusters.

Pour bien comprendre le concept de clustering par densité, nous présentons un exemple de cette méthode (Figure 2-4). Si le nombre de voisins d'un objet dépasse un seuil minimum (dans l'exemple MinPts >=3) alors l'objet est appelé noyau (centre du cluster). o, r, p, s, m, q sont considérés comme des noyaux. Les voisins d'un objet sont à des distances inférieures au rayon qui est une distance de voisinage définie par l'utilisateur. q est considéré comme directement accessible par densité à partir de m et indirectement accessible à partir de p mais p n'est pas indirectement accessible à partir de q car ce n'est pas un noyau (nombre de voisins < MinPts). Cette accessibilité est appelée Density-reachable. o, r, s forment une connexion ou une liaison dite de densité (Density-connected) forme le cluster.

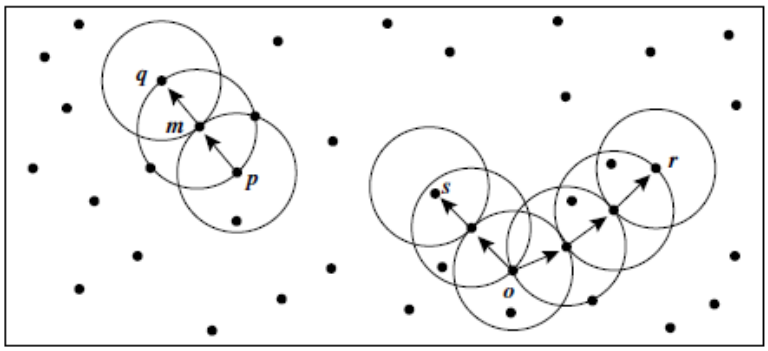


Figure 2-4 : Exemple de construction de clusters à base de densité
(Miller & Han 2009)[50].

Après avoir présenté les concepts de base du clustering de densité à partir d'un exemple, nous détaillerons l'un des algorithmes les plus utilisé, à savoir DBSCAN. Cet algorithme a été présenté pour la première fois lors de la conférence KDD en 1996 par Ester (Ester et al., 1996). L'algorithme commence par sélectionner arbitrairement un point p parmi l'ensemble des points, trouve tous les points qui lui sont accessibles à partir d'une distance radius (Eps). Si p a un nombre de voisins dépassant MinPts alors il est considéré comme noyau et un cluster ayant p comme centre est créé. Si p est un point frontière et il y a pas de point qui lui sont accessibles alors le point suivant est visité. Le processus continue jusqu'à ce que tous les points soient traités. Pour plus de détails sur DBSCAN, l'algorithme est présenté dans l'Algorithme 2-3.

Algorithme DBSCAN

```
Entrée : D (les données), eps (radius ou distance de voisinage), MinPts
(minimum d'objets voisins pour constituer une classe)
Sortie : C (les classes)
Début

C = 0
pour chaque point P non visité des données D
marquer P comme visité
PtsVoisins = epsilonVoisinage(P, eps)
si tailleDe(PtsVoisins) < MinPts
mark P as NOISE
sinon
C++
expandCluster(D, P, PtsVoisins, C, eps, MinPts)
```

[50] Page 419

```
Fin
expandCluster(D, P, PtsVoisins, C, eps, MinPts)
ajouter P au cluster C
pour chaque point P' de PtsVoisins
si P' n'a pas été visité
marquer P' comme visité
PtsVoisins' = epsilonVoisinage(D, P', eps)
si tailleDe(PtsVoisins') >= MinPts
PtsVoisins = PtsVoisins U PtsVoisins'
si P' n'est membre d'aucun cluster
ajouter P' au cluster C

epsilonVoisinage(D, P, eps)
retourner tous les points de D qui sont à une distance inférieure à
epsilon de P
```

Algorithme 2-3 : Algorithme DBSCAN.

2.2.1.4. Les Séries chronologiques

Les Séries chronologiques permettent d'identifier les séquences similaires à une portion de données, de prévoir et de déterminer les causalités à partir des bases de données de séries temporelles. Ces bases de données peuvent êtres des bases de données séquentielles (voir section 2.2.1.1.2) ou de valeurs (mesures) obtenues pour des intervalles de temps, comme le cas des mesures de température.

Les bases de données séquentielles ne sont pas forcément des bases de données de séries temporelles car elles peuvent contenir des événements séquentiels non étiquetés dans le temps.

L'objectif des Séries temporelles est de chercher dans de grandes quantités de données, des motifs similaires, réguliers, cycliques, les comportements (tendance), les impulsions, etc. Cela permet de modéliser le mouvement en le décomposant en une série de mouvements basiques (tendance, saisonnalité, cyclicité, irrégularités, horizontalité, etc.) et faire des prédictions de mouvements futures. La tendance correspond à une croissance ou une décroissance de la variable avec le temps, la saisonnalité est caractérisée par une série qui change selon un facteur saisonnier (mois, jour de la semaine, saison), la cyclicité est analogue à une loi saisonnière mais la longueur du cycle est supérieure à un an et ne se répète pas nécessairement à des intervalles de temps réguliers, l'irrégularité est un comportement qui présente une variabilité par rapport au comportement global et enfin l'horizontalité correspond à une stationnarité de la série où les données ne représentent aucune tendance. C'est dans un objectif de prise de décision

et de prédiction des comportements d'un système au cours du temps que les séries temporelles sont décomposées, analysées, expliquées et modélisées.

Plusieurs applications utilisent les séries temporelles comme le marketing (voir Figure 2-5), la médecine, la chimie et les finances.

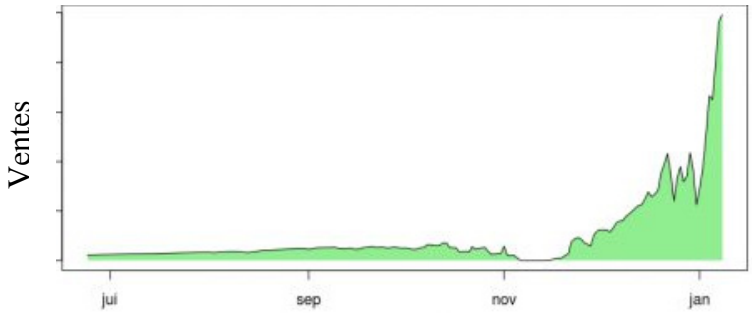


Figure 2-5 : Exemple d'une série chronologique représentant des bénéfices de ventes.

2.2.1.5. Analyse d'aberrations

Les aberrations, appelées aussi outliers (terme anglais) sont des données qui ne suivent pas le comportement ou le modèle général. Les outliers doivent être manipulés avec prudence car ils peuvent décrire une erreur ou une variabilité importante dans le comportement du système étudié. Ces outliers sont soit supprimés pour qu'ils n'influencent pas les tendances globales, ou au contraire mis en évidence s'ils véhiculent une information utile à un domaine d'application particulier. Il y a énormément de domaines qui s'intéressent à ce problème de détection d'aberrations.

L'œil humain est très efficace pour la détection d'outliers à partir de visualisations s'il n'y a pas énormément de données affichées. Mais dès que la charge visuelle augmente et les contraintes de temps se manifestent, l'œil atteint ses limites (Han & Kamber 2006). Ainsi il y a un réel besoin d'automatiser cette détection. Dans cet objectif, différentes méthodes informatiques ont été proposées dans la littérature. Ces méthodes utilisent différentes approches comme les statistiques, les mesures de distances, la déviation et la densité (Han & Kamber 2006)[51]. Chacune de ces approches va être détaillée dans les sous-sections ci-après.

[51] Page 452

2.2.1.5.1. Détection d'outliers par les statistiques

Plusieurs méthodes statistiques pour la détection des valeurs aberrantes existent que ce soit sur des données univariées ou multivariées. Pour les méthodes univariées, nous pouvons citer la plus naïve qui consiste à définir un seuil à partir duquel les valeurs sont considérées comme outliers, les méthodes algébriques qui calculent les distances qui séparent les valeurs de la distribution de son centre (moyenne, médiane, quantiles, etc.), les méthodes graphiques ou visuelles comme les Box Plot (Cf. section 1.5.1) (Vasyechko et al., 2005).

Il peut s'avérer parfois utile d'analyser plusieurs variables (multivarié) pour détecter les outliers. Dans certaines variables, la détermination univariée de valeurs aberrantes peut aboutir à une mauvaise classification (Vasyechko et al., 2005).

2.2.1.5.2. Détection d'outliers par mesure

Les outliers ici sont des données qui n'ont pas assez de voisins. Les voisins sont des données se trouvant à un certain seuil de distance comme le radius de DBSCAN (Cf. section 2.2.1.3). La distance peut être métrique comme la distance euclidienne ou non-métrique comme le coût de déplacement (prix, temps, consommation de carburant, etc.). Plusieurs algorithmes de détection d'outliers basés sur la distance existent comme Index-based algorithm, Nested-loop et Cell-based algorithm.

La détection des outliers par les distances et les statistiques posent quelques difficultés quant à l'analyse de données non uniformément distribuées et/ou ayant des densités variables (Figure 2-6). L'exemple donné pas J. Han (Han & Kamber 2006)[52], illustre bien la différence entre les résultats des différents types de méthodes. Il représente la distribution de valeurs par croisement de deux variables non spatiales.

[52] Page 456

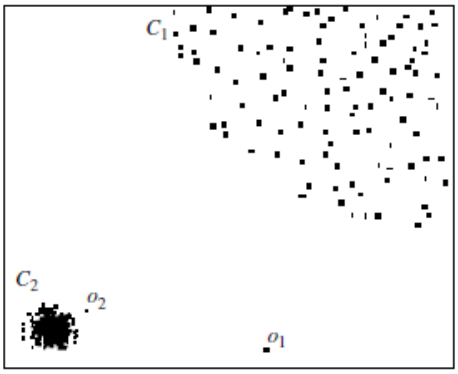


Figure 2-6 : Détection des outliers. Il est évident que o1 et o2 sont des outliers mais o2 ne sera pas considéré comme outliers dans la méthode basée sur les statistiques et les distances car o2 est dans le voisinage des objets de C2. La distance qui sépare o2 de c2 est petite
(Han & Kamber 2006).

2.2.1.5.3. Détection d'outliers par déviation

La détection des outliers basée sur la déviation (Zhang & Feng 2009) se focalise sur l'identification des caractéristiques principales des groupes dans l'objectif de découvrir les objets outliers qui sont déviés de ces caractéristiques des groupes (Han & Kamber 2006). Cette méthode n'utilise ni les distances, ni les tests statistiques.

2.2.1.5.4. Détection d'outliers basée sur les densités

Les outliers sont considérés dans ce paragraphe comme des petits groupes d'objet ou de données non denses et isolés. L'approche Local Outlier Factor (LOF) (Breunig et al., 2000) est souvent utilisée pour ce type de détection.

Le clustering est utilisé aussi pour ce type de détection en identifiant les objets isolés qui ne font partie d'aucun cluster (Han & Kamber 2006)[53].

2.2.2. La fouille de données spatiales

Selon ESRI[54], 80% des données possèdent une composante spatiale[55]. La prise en compte de cette composante en explorant les données et les relations spatiales par la

[53] Chapitre 7, page 383
[54] Editeur de système de gestion de l'information géographique
[55] http://www.esri.com/library/whitepapers/pdfs/reveal-more-value.pdf

fouille de données, ouvre des perspectives intéressantes pour la découverte de nouveaux modèles. En effet, la simple visualisation des données sur une cartographie peut permettre la découverte de connaissances comme l'avait démontré le Dr. John Snow (Berche 2007)[56] en trouvant des corrélations entre la localisation des cas de choléra qui avait touché une région du nord de Londres en 1854 et la localisation des puits d'eau. Le Dr. J. Snow est considéré comme l'initiateur de l'analyse cartographique. Cette découverte visuelle des connaissances peut être utilisée sur un volume de données limité mais elle doit être automatisée pour de grandes quantités de données. Cette automatisation est supportée par la fouille de données spatiales.

Le domaine de la fouille de données spatiale est un domaine à part entière qui a été amorcé par les premiers travaux de Koperski, J. Han (Han et al., 1997) et Ester (Ester et al., 1997). Ce domaine s'intéresse à la découverte de modèles dans une base de données spatiales (Committee 2003). La principale caractéristique de ce domaine est sa prise en compte de la dimension spatiale et des relations entre les objets (Chelghoum & Zeitouni 2004). Les objets étudiés sont des thèmes rassemblant les objets de même type. Ces thèmes ne sont rien d'autre que des tables avec un attribut de localisation où les interactions entre les objets sont représentées par des prédicats et des tables de distances.

Les techniques de fouille de données classiques ne peuvent pas supporter directement la composante spatiale dans l'exploration automatique des données pour trois raisons :

- Les données géospatiales sont plus complexes et contiennent des objets, des nombres et des catégories (hétérogénéité). Les objets ont différents types de données, plusieurs types de géométries possibles (point, lignes, polygone) et les relations entre ces objets sont souvent implicites (relations topologiques, relations de distance, direction). Les relations topologiques peuvent être entre toutes les combinaisons possibles de géométries (points*points, point*ligne, etc.).

- Les techniques classiques utilisent des données explicites mais dans le géospatial, les données peuvent être construites à la volée (par exemple l'Overlaping) et les relations spatiales sont la plupart du temps implicites.

[56] Page 70

- Les données de fouille de données classique sont traitées indépendamment des autres (on suppose qu'elles sont indépendantes) alors que dans la fouille de données spatiales, les données sont spatialement dépendantes les unes des autres.

Les méthodes de fouille de données classiques doivent alors être adaptées pour supporter la composante spatiale. Des travaux s'intéressent à l'adaptation de ces méthodes en transformant par exemple des problèmes de la fouille de données spatiales en des problèmes de fouille de données multi-tables (Chelghoum et al., 2006). Contrairement à une représentation *individu-enregistrement* qui est souvent utilisée dans la fouille de données classique, la fouille de données multi-tables se base sur une représentation des données sous forme d'individus dans une table associée potentiellement à plusieurs enregistrements dans des tables secondaires (Dhafer et al., 2012). Dans la littérature, parmi les méthodes classiques principalement adaptées, nous trouvons, les règles d'association, les statistiques et la classification. Les méthodes de fouille de données spatiales sont la plupart du temps scindées en deux types (Aufaure et al., 2000) :

- **Les méthodes exploratoires ou mono thématiques** :

Ces méthodes s'appliquent à un seul thème géographique et permettent d'identifier les écarts et/ou les similarités entre les objets. Ces méthodes sont souvent basées sur les analyses statistiques et elles sont scindées à leur tour en trois familles :

1. Description synthétique : autocorrélation, généralisation, densité et lissage.
2. Spécificités locales : autocorrélation locale, analyse factorielle locale
3. Groupement de données : Clusters de points, clusters de trajectoires, etc.

- **Les méthodes décisionnelles ou multithématiques :**

Ces méthodes s'appliquent à plusieurs thèmes géographiques dans le but d'expliquer les écarts et les caractéristiques des groupements. Donc elles peuvent être vues comme une deuxième phase après l'exploration. Parmi ces méthodes nous pouvons citer la caractérisation, les règles d'association et la classification spatiale adaptées à la fouille de données spatiales. Ces méthodes sont souvent basées sur de l'induction et les techniques de bases de données spatiales.

La section 2.2.1.3 explique l'intérêt des méthodes de groupement par densité par rapport aux autres. Le groupement de données spatiales, et plus particulièrement, le groupement de positions permet d'organiser les positions en classes d'une façon à ce que les objets similaires soient dans les mêmes classes et les objets dissimilaires soient dans des classes différentes. Les méthodes de groupement basées sur la densité séparent les espaces denses (les clusters) des espaces moins denses (le bruit) (Figure 2-7). Le cluster de densité a été défini par J. Han comme étant un ensemble maximal de points connectés par la densité (Han & Kamber 2006)[57].

Figure 2-7 : Exemple de clusters de densité où chaque cluster a une grande densité de points (G. Gasso & P.Leray, 2013)[58].

2.2.3. La fouille de données d'objets mobiles

De plus en plus d'objets mobiles sont équipés de capteurs générant un nombre important de positions qui sont reçues, stockées et visualisées dans le but de surveiller l'évolution des systèmes (Etienne et al., 2010). L'application de la fouille de données aux historiques de positions d'objets mobiles ouvre de nouvelles perspectives intéressantes. En effet, cela va permettre d'identifier les comportements normaux et anormaux d'objets en mouvement et faire des prédictions ou des classements. D'un autre point de vue, le déplacement pose des problèmes aux modèles de données géospatiales et à la fouille de données spatiales. En effet, il est difficile d'identifier et de catégoriser les objets mobiles sur des trajectoires (Committee 2003).

Plusieurs travaux de recherche se sont intéressés à la fouille de données d'objets en mouvement. Parmi ces travaux nous pouvons citer les travaux de l'équipe de J. Han (Lee

[57] Page 418
[58] Gilles Gasso et Phillippe Leray, Clustering, INSA Rouen, 2013

et al., 2008a; Li et al., 2010c), de K. Zeitouni (Kharrat et al., 2009) et Giannotti et Nanni (Giannotti et al., 2007). Dans la littérature, des travaux font la distinction entre la fouille de patterns d'objets mobiles et la fouille de données de trajectoires. On considère dans la suite de ce travail que la fouille de données d'objets mobiles regroupe les méthodes qui prennent en compte la dimension temporelle et celles qui ne s'intéressent qu'à la dimension spatiale.

La distinction qui nous captive est celle qui considère deux types de méthodes de fouille de données : le premier type voit les objets mobiles comme des objets qui se déplacent dans un espace ouvert et le second, comme des objets se déplaçant dans un espace soumis à des contraintes de réseaux.

2.2.3.1. Espace d'évolution ouvert

J. Han propose des méthodes de fouille de données d'objets mobiles pour révéler des mouvements collectifs, des clusters de trajectoires (Lee et al., 2007), détecter des trajectoires aberrantes (Lee et al., 2008a), détecter des périodicités dans les déplacements (Li et al., 2010b) et faire des classification d'objets selon leur trajectoire (Lee et al., 2009). Avant Li, Cao (Cao et al., 2007) s'est penché sur la question des périodicités et a proposé une méthode pour les découvrir. Parmi les méthodes de clustering de trajectoires, on trouve la détection des ensembles d'objets se déplaçant en convois comme la méthode proposé par l'équipe de Jeung (Jeung et al., 2008a).

2.2.3.1.1. Détection de trajectoires aberrantes

Les trajectoires aberrantes représentent des mouvements qui ne ressemblent pas au comportement général. Les trajectoires aberrantes sont de deux types : les trajectoires qui ne suivent pas les mêmes chemins que les autres objets, et les sous-trajectoires qui ne suivent pas la même tendance que les autres sous-trajectoires se trouvant dans son voisinage (voir la sous-trajectoire aberrante TR3 de la Figure 2-8). Les sous-trajectoires peuvent contenir une ou plusieurs partitions, la partition étant l'unité la plus petite d'une trajectoire après la position.

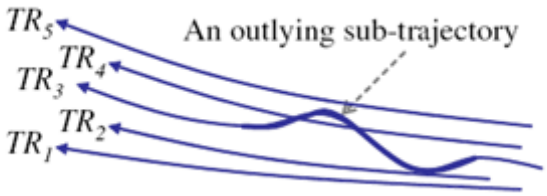


Figure 2-8 : Exemple d'une sous-trajectoire aberrante (Lee et al., 2008a).

La plupart des algorithmes proposés dans la littérature pour la comparaison de trajectoires considèrent des trajectoires complètes sans les partitionner en sous trajectoires. Cette approche empêche la détection de partitions de trajectoires outliers (Figure 2-8). Lee (Lee et al., 2008a) présente un algorithme de détection de trajectoires et de sous-trajectoires aberrantes appelé TRAOD (TRAjectory Outlier Detection). L'algorithme travaille en deux phases, il commence par une phase de partitionnement des trajectoires en sous trajectoires puis il enchaine par une deuxième phase de détection des outliers basée sur une mesure de distance et de densité (Figure 2-9). L'objectif est de pouvoir détecter les trajectoires et les sous trajectoires outliers dans les partitions denses et non denses.

Figure 2-9 : Illustration du fonctionnement de l'algorithme TRAOD.

Comment sont effectués les partitionnements des trajectoires ?

L'algorithme prend en entrée un ensemble de trajectoires sous forme de séquences de positions successives. Le partitionnement de ces trajectoires est basé sur deux approches différentes : une approche basée sur le principe MDL (*Minimum Description Length*), et une approche à deux niveaux.

Le principe MDL (Grünwald et al., 2005) est souvent utilisé dans la théorie de l'information. Le problème de partitionnement est ainsi transformé en un problème MDL. L'intérêt de ce partitionnement est son aptitude à optimiser les données sans avoir besoin de paramètres en entrée comme le cas de la réduction de données spatio-temporelles (Cao et al., 2006). Le partitionnement basé sur MDL cherche à trouver le meilleur compromis entre la longueur de la description de l'hypothèse L(H) et la longueur de la description des données lorsqu'elles sont codées à l'aide de l'hypothèse L(D|H). Le compromis est une fonction qui minimise la somme des deux éléments L(H) et L(D|H). H désigne l'hypothèse et D les données.

Dans l'algorithme TRAOD, l'auteur essaye de minimiser la longueur de la nouvelle trajectoire simplifiée en minimisant au maximum le nombre de positions de la trajectoire (concision) et en respectant le plus possible la représentation de la trajectoire originale (précision). La nouvelle trajectoire simplifiée, doit donc assurer une meilleure représentation de la trajectoire originale à travers ses partitions, et une meilleure concision à travers le nombre de partitions constituant cette nouvelle trajectoire.

Plus il y a de partitions, meilleure sont la représentation et la qualité de détection, mais les performances sont réduites. La complexité algorithmique de TRAOD est de $O(n^2)$, telle que n est le nombre total de partitions des trajectoires.

Dans l'objectif de réduire l'espace de recherche et avoir de meilleures performances, une approche de partitionnement en deux niveaux est proposée : des partitions grossières en premier niveau, puis ; si nécessaire, les partitions grossières sont partitionnées à leur tour, en partitions fines en deuxième niveau (Figure 2-10).

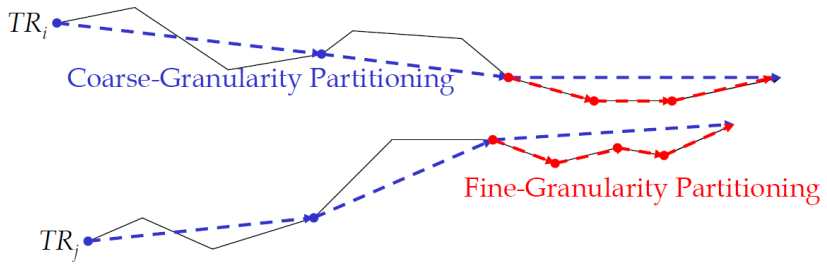
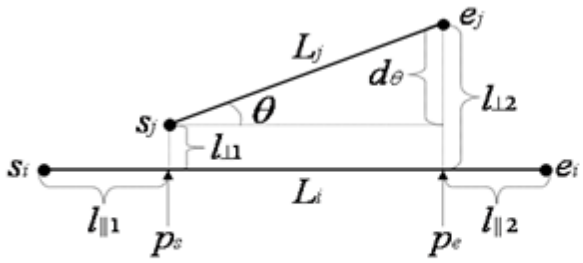


Figure 2-10 : Partitionnement de trajectoires en deux niveaux, t-partions grossières et t-partition fines (Lee et al., 2008a).

L'algorithme TRAOD inspecte les partitions grossières pour partitionner en fines partitions celles qui sont susceptibles d'être outliers. L'identification des partitions pouvant être des outliers est effectuée en se basant sur le concept de borne supérieure (*ub)* et de borne inferieure (*lb)* de la distance entre deux partitions grossières Li et Lj.

Le partitionnement en fines partions n'est pas nécessaire si la borne inférieure est supérieure à la distance entre Li et Lj ($lb > D$). Le partitionnement en fines partitions est par contre effectif si la borne supérieure est inférieure ou égale à la distance entre Li et Lj ($ub \leq D$).

La distance entre deux partions est calculée sur la base d'une distance perpendiculaire $d_{\perp}$, parallèle $d_{\parallel}$ et angulaire d_{θ} (Figure2-11) qui sont pondérées par rapport à des poids $w(w_{\perp}, w_{\parallel}, w_{\theta})$ pour donner plus de poids, score ou d'importance à une distance par rapport à une autre. Par exemple, dans l'une des expérimentations menées par Lee et son équipe (Lee et al., 2008a), le poids de la distance angulaire est fixé à cinq fois la distance parallèle et la distance perpendiculaire pour la découverte de trajectoires aberrantes dans le déplacement d'ouragans.



$$d_{\perp} = \frac{l_{\perp 1}^2 + l_{\perp 2}^2}{l_{\perp 1} + l_{\perp 2}}$$
$$d_{\parallel} = \mathrm{MIN}(l_{\parallel 1}, l_{\parallel 2})$$
$$d_{\theta} = \|L_j\| \times \sin(\theta)$$

$$dist(L_i, L_j) = w_{\perp} d_{\perp} + w_{\parallel} d_{\parallel} + w_{\theta} d_{\theta}$$

Figure 2-11 : Calcul de distance entre deux partitions de trajectoires intégrant la distance perpendiculaire, parallèle et angulaire (Lee et al., 2008a).

Le *lb* et *ub* sont calculés comme suit :

$$lb(L_i, L_j, dist) = w_{\perp} \cdot lb(Li, Lj, d_{\perp}) + w_{\parallel} \cdot lb(Li, Lj, d_{\parallel}) + w_{\theta} \cdot lb(Li, Lj, d_{\theta})$$
$$ub(L_i, L_j, dist) = w_{\perp} \cdot ub(Li, Lj, d_{\perp}) + w_{\parallel} \cdot ub(Li, Lj, d_{\parallel}) + w_{\theta} \cdot ub(Li, Lj, d_{\theta})$$

Le calcul des bornes supérieures (*ub*) et inférieures (*lb*) des distances perpendiculaires, parallèles et angulaires sont présentées ci-après :

- **Distance perpendiculaire** :

$$lb(L_i, L_j, d_\perp) = MIN\ (l_{\perp 1}, l_{\perp 2}) - (\max l_\perp(L_i) + \max l_\perp(L_j))$$
$$ub(L_i, L_j, d_\perp) = MAX\ (l_{\perp 1}, l_{\perp 2}) - (\max l_\perp(L_i) + \max l_\perp(L_j))$$

Telle que, $l_{\perp 1}$ et $l_{\perp 2}$ sont des distances euclidiennes des deux extrémités de la partition L_j sur leur projection sur L_i,

$\max l_\perp(L_i)$: Distance perpendiculaire maximale entre une partition grossière et ses partitions fines,

$\max l_\theta(L_i)$: Angle maximal entre une partition grossière et ses partitions fines.

La figure suivante (Figure 2-12) représente les différentes composantes de la formule définie ci-dessus et montre comment ils sont calculées.

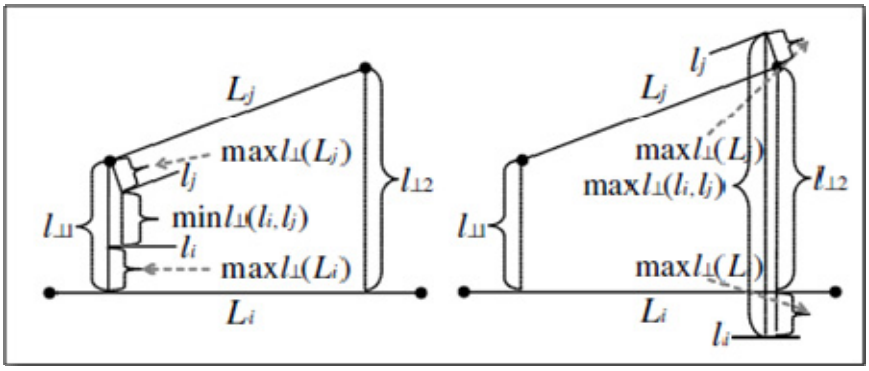


Figure 2-12 : Illustration des distances qui rentrent dans le calcul des composantes *lb* et *ub* de la distance perpendiculaire (Lee et al., 2008a).

- **Distance parallèle :**

$lb(L_i \text{ et } L_j, d_{//}) = 0$ *si les partitions L_i et L_j sont superposées ou jointes*
$= d_{//}(L_i, L_j)$ *si elles sont disjointes.*

$ub(L_i \text{ et } L_j, d_{//}) = MAX(len(L_i), len(L_i))$ *si les partitions L_i et L_j sont jointes*
$= len(L_i)+len(L_j)-d_{//}(L_i, L_j)$ *si elles sont superposées*
$= len(L_i)+len(L_j)+d_{//}(L_i, L_j)$ *si elles sont disjointes*

- **Distance angulaire :**

$$lb(L_i, L_j, d_\theta) = MIN(\min len(L_i), \min len(L_j)) * \sin(\theta - \max\theta(L_i) - \max\theta(L_j));$$
$$ub(L_i, L_j, d_\theta) = MIN(\max len(L_i), \max len(L_j)) * \sin(\theta + \max\theta(L_i) + \max\theta(L_j));$$

Telle que θ est l'angle entre 2 partitions grossières L_i et L_j et $max\theta(L_i)$ est l'angle maximal entre la partition grossière L_i et ses partitions fines.

Comment s'effectue la détection ?

L'algorithme prend en entrée trois paramètres qui sont la distance entre les partitions de trajectoires voisines (D), la proportion de trajectoires voisines pour ne pas être une partition aberrante (p) et la proportion de la longueur des outliers dans une trajectoire pour qu'elle soit aberrante (F). La détection des outliers est donc basée sur un seuil de distance (D) et un seuil de densité (p). L'algorithme compare la distance entre les partitions des trajectoires et la distance fournie par l'utilisateur (D). L'algorithme TRAOD est présenté ci-après (Algorithme 2-4).

- Une partition est dite outlier si elle n'a pas assez de partitions voisines similaires (voir Figure 2-13)

Nombre trajectoires voisines $\leq$ *(1-p)* nombre total de trajectoires,*

- Une trajectoire est dite outlier si elle contient une proportion de longueur suffisante de partitions outliers :

$$\frac{\text{la somme des longueurs des partitions outliers de la trajectoire i}}{\text{la somme des longueurs de toutes les partitions de la trajectoire i}} \geq F$$

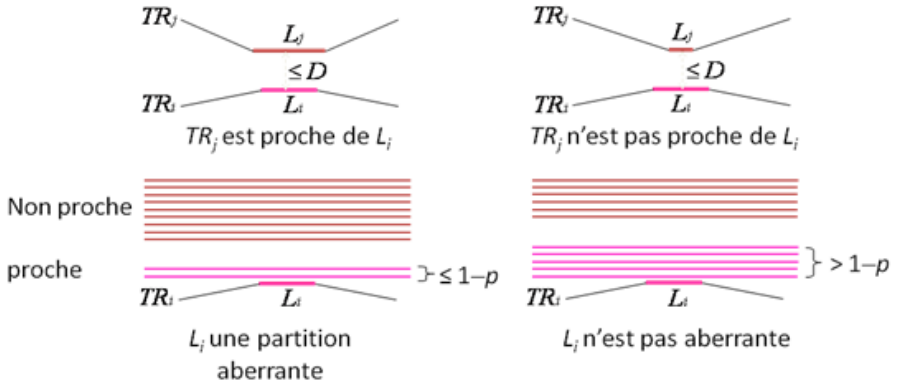


Figure 2-13 : Détection des outliers basée sur la distance et la densité (source Lee[59]).

[59] http://dm.kaist.ac.kr/jaegil/slides/icde08.ppt

Cette approche ne prend pas en compte l'hétérogénéité de la distribution des trajectoires : les partitions se trouvant dans des régions denses ont relativement un nombre de partitions voisines plus grand que celles se trouvant dans des régions éparses et ont donc moins de chances d'être détectées comme étant outliers. Un facteur d'ajustement $adj(L_i)$ est multiplié par le nombre de trajectoires voisines pour ajuster la densité locale par rapport à la densité moyenne. Ce facteur est un rapport entre la moyenne de toutes les densités des partitions et la densité de la partition L_i.

Algorithme TRAOD (TRAjectory Outlier Detection)

Entrée : I = {TR1, …, TRn} (ensemble de trajectoires)
Sortie : O = {O1, …, Om} (ensemble de trajectoires outlier avec leurs partitions associées)
Début
/* Partitioning Phase */
Pour chaque TR ∈ I
Partitionner TR en un ensemble de segments de lignes L
Mettre L dans un ensemble D
/* Detection Phase */
Pour chaque P ∈ D
Vérifier si P est une partition outlier
Pour chaque TR ∈ I
Retourner les TR ayant suffisamment d'outlier
Fin

Algorithme 2-4 : Algorithme TRAOD.

2.2.3.1.2. Clustering de trajectoires

Le clustering de trajectoires s'inspire du clustering de positions (Cf. section 2.2.2) où les positions sont remplacées par des trajectoires. L'objectif de ce clustering est de regrouper les trajectoires individuelles en des trajectoires représentatives du mouvement général. Deux types d'algorithmes existent : ceux qui traitent les trajectoires dans leur ensemble comme ceux proposés par Gaffney (Gaffney et al., 2006), et l'algorithme TRACLUS de L. Lee (Lee et al., 2007) qui propose de partitionner les trajectoires en segments de lignes avant de regrouper les partitions similaires. L'algorithme TRACLUS (TRAjectory CLUStering) permet de faire du clustering en deux phases, une phase de partitionnement et une phase de groupement (voir Figure 2-14). Il commence par une phase de partitionnement des trajectoires en petites partitions, basée sur un algorithme

glouton[60] appelé *Approximate Trajectory Partitioning*. Cet algorithme est proposé par le même auteur de TRACLUS pour le partitionnement de trajectoires car le partitionnement par MDL coûte cher en performance. Une approche considérant que les optimaux locaux peuvent être des optimaux globaux va alors réduire le temps de partitionnement. L'algorithme glouton de partitionnement proposé reste basé sur le problème MDL vu dans la section précédente. Les partitions résultantes de la simplification des trajectoires par *Approximate Trajectory Partitioning* sont utilisées dans la deuxième phase pour grouper celles qui sont similaires. Le groupement des partitions de trajectoires est basé sur une extension de l'algorithme DBSCAN.

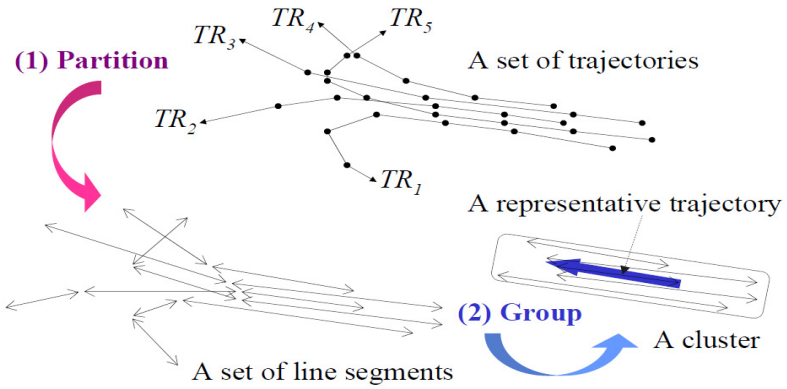


Figure 2-14 : Les deux phases de l'algorithme TRACLUS : Partition et groupement des partions pour la découvertes des clusters de trajectoires (Lee et al., 2007).

Comme DBSCAN, TRACLUS utilise un paramètre de distance de voisinage (ε-voisinage) tel que les partitions se trouvant à une distance inférieure de ε-voisinage sont des partitions voisines ($N\varepsilon(Li) = \{Lj \in D \mid dist(Li, Lj) \leq \varepsilon\}$). Si le nombre de partitions voisines d'une partition Li dépasse le nombre minimal de partitions ($|N\varepsilon(Li)| \geq MinLns$) alors c'est une partition noyau. Une trajectoire représentative est donc créée pour chaque cluster de trajectoires (Trajectoire rouge sur la Figure 2-15) à partir de ces partitions noyaux qui sont les segments principaux des ensembles de partitions voisines. Des points sont créés pour chaque début et fin d'une partition et le cluster est créé en suivant une

[60] Appelé greedy algorithm en anglais, c'est un algorithme approximatif qui essaye d'approcher un optimum global en choisissant des optimaux locaux à chaque itération, Thierry Mautor, PRISM, UVSQ, 2009.

ligne de balayage. La direction est calculée comme une moyenne des vecteurs de direction des partitions voisines.

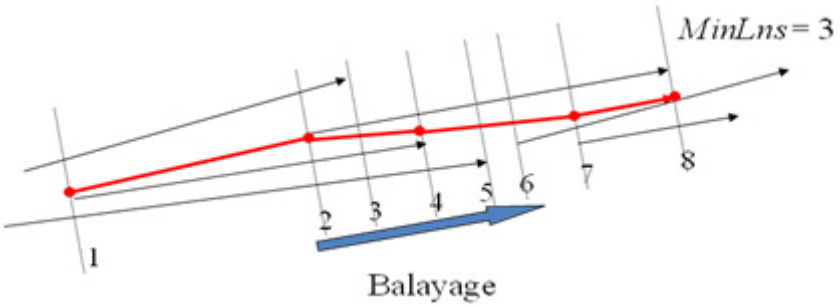


Figure 2-15 : Exemple de construction d'un cluster de trajectoires pour MinLns =3.

Les définitions de DBSCAN ont été adaptées pour TRACLUS. Une partition Li est dite directement joignable par densité à partir de Lj par rapport à ε-voisinage et MinLns, si Li ∈ Nε(Lj) et |Nε(Lj)| ≥ MinLns et les partitions sont joignables par densité avec la propriété de fermeture transitive de la jointure directe de densité. Li est connecté à Lj par la densité, s'il existe Lk tel que Li et Lj sont joignables par densité de Lk. Les ensembles connectés par densité forment les clusters de trajectoires. L'algorithme de TRACLUS est présenté ci-après (Algorithme 2-5).

Algorithme TRACLUS

Entrée : I (ensemble de trajectoires) {TR1,…,TRnum_traj}
Sortie : O (ensemble de clusters) {C1,…, Cnum_clus},
R (ensemble représentative des trajectoires)
Début

/*Phase de partitionnement*/
Pour chaque TRi ∈ I
Partitionnement des trajectoires avec une méthode approximative
Obtenir un ensemble L de segments à partir des résultats du partitionnement
Mettre L dans D

/*Phase de groupement*/
Clustering des segments de trajectoires de D
Obtenir un ensemble O de clusters à partir des résultats de clustering
Pour chaque C ∈ O
Construire les trajectoires représentatives
Obtenir un ensemble R des trajectoires représentatives à partir des résultats
Fin

Algorithme 2-5 : Algorithme TRACLUS.

Il est à remarquer que la trajectoire représentative du clustering est sensible à l'hétérogénéité des partitions de trajectoires : directions différentes et décalage non perpendiculaire entre les débuts et les fins des partitions. La Figure 2-16 illustre bien l'effet de ce décalage sur la forme de la trajectoire représentative.

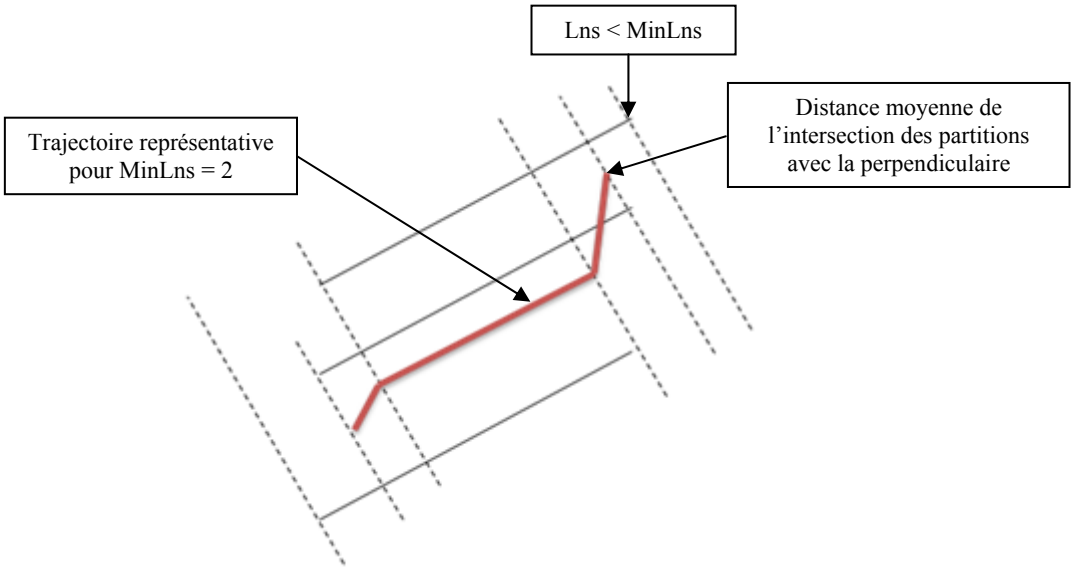


Figure 2-16 : Illustration du calcul de la trajectoire représentative à partir de trois partitions parallèles avec des départs décalés.

2.2.3.1.3. Classement de trajectoires

Comme vu dans la section 2.2.1.2, le classement permet de modéliser les classes caractéristiques des groupes de données pour permettre le classement de nouvelles données. Le classement de trajectoires est une extension de ce problème aux données mobiles. Parmi les algorithmes proposés dans la littérature pour le classement des trajectoires, nous pouvons citer *Trajectory Classification Using Hierarchical Region-Based and Trajectory-Based Clustering* (Lee et al., 2008b). TRACLASS pose le postulat que les objets similaires fréquentent les mêmes régions et ont les mêmes comportements.

Cet algorithme s'exécute en trois phases, il commence par une phase de partitionnement de trajectoires, puis de groupement et ensuite de classement. Après le partitionnement de trajectoires, il trouve les régions rectangulaires majoritaires où le nombre des partitions de la même classe dans une région est supérieur ou égal à ψ (Figure 2-17- (1) (2)). Une fois les régions denses identifiées (groupement basé sur les régions), les trajectoires se trouvant à l'extérieur des régions sont partitionnées et regroupées (groupement basé sur les trajectoires) en prenant en compte les régions

trouvées au préalable (Figure 2-17- (3) (4)). Deux groupes sont dits connectables s'ils partagent suffisamment de trajectoires (fraction des trajectoires ≥ χ). Le groupement est donc un groupement en deux niveaux, basé sur les régions et basé sur les trajectoires.

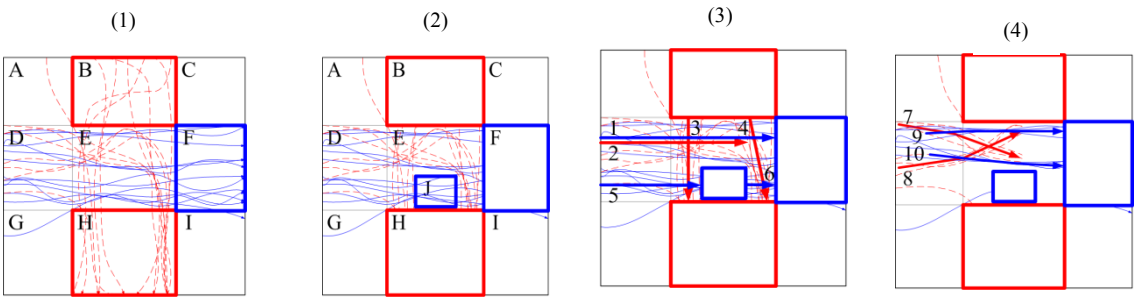


Figure 2-17 : Exemple de groupement de régions (1) (2) puis de trajectoires (3) (4) (Lee et al., 2008b).

Un cluster basé trajectoires est un ensemble de partitions qui sont connectées par densité et qui appartiennent à la même classe comme illustré sur la Figure 2-18.

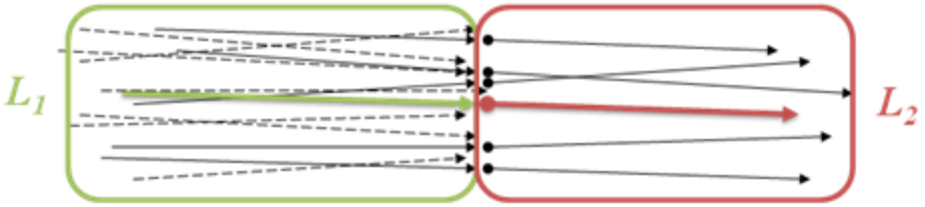


Figure 2-18 : Le groupement de partitions de trajectoires selon le libellé de chaque classe (Lee et al., 2008b).

Comme il est représenté dans le processus de classification de TRACLASS (Figure 2-19), après la phase de groupement basé sur les régions et sur les trajectoires, le groupement basé sur les deux est effectué. La phase de groupement régions s'arrête soit quand il n'y a plus de régions hétérogènes, soit quand la taille minimale d'une région est atteinte. En ce qui concerne le groupement basé trajectoires, il s'arrête quand tous les groupes sont détectés. Le classifieur mis en œuvre à la fin du processus de classification de TRACASS peut être utilisé pour classer de nouveaux comportements.

L'idée de la phase de classification consiste à convertir chaque trajectoire en un vecteur caractéristique dont chaque entrée est soit un cluster basé sur les régions, sur les trajectoires soit une caractéristique numérique. Après avoir défini ces vecteurs caractéristiques, le classificateur sera construit par une méthode de généralisation des

classificateurs linéaires appelée *Support Vector Machines* (SVM). Le SVM « *Support Vector Machines* » est un ensemble de techniques d'apprentissage supervisé dans des problèmes de discrimination et de régression.

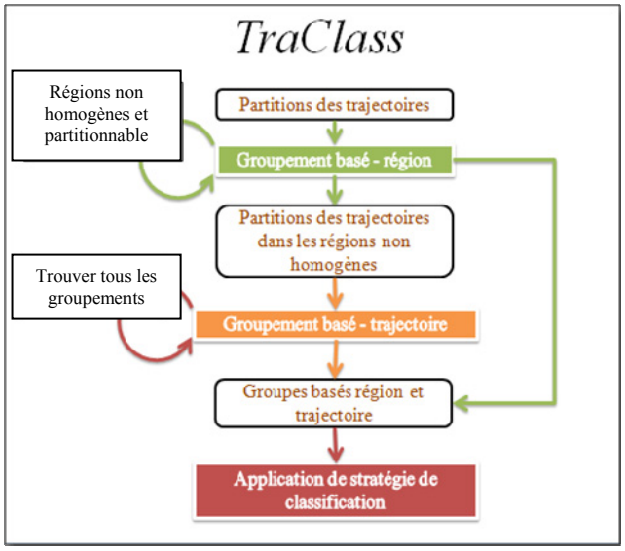


Figure 2-19 : Processus de classement de l'algorithme TRACLASS (Lee et al., 2008b).

Prenons un exemple d'application de cet algorithme sur un ensemble de trajectoires de navires évoluant entre un port A et un port B (voir Figure 2-20) qui a permis la construction d'un classifieur de trajectoires.

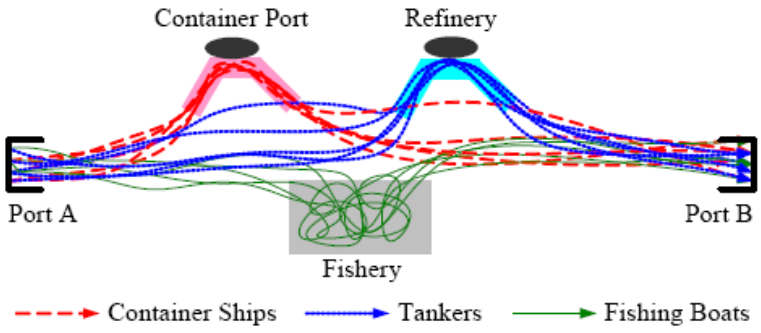


Figure 2-20 : Exemple de classement de trajectoires de navires (Lee et al., 2008b).

L'algorithme détecte d'abord les régions ayant le plus de trajectoires homogènes (régions autour de Container Port, Rafinery et Fishery), puis découvre les groupes de trajectoires partagées entre les régions (mouvements communs des classes de trajectoires). Les régions et mouvements communs de chaque classe de trajectoire sont utilisés comme caractéristiques de trajectoires pour discriminer les types des navires

selon leur trajectoire. Le classifieur construit à partir des vecteurs caractéristiques des trajectoires de navires va permettre d'associer à chaque nouvelle trajectoire l'une des classes suivantes : Container Ships, Tankers et Fishing Boats.

2.2.3.1.4. Détection de Clusters d'objets mobiles

Les clusters d'objets mobiles ou de mouvements permettent de détecter des objets qui se déplacent conjointement pendant un intervalle de temps long (Kalnis et al., 2005). Ce sont des clusters spatiaux successifs dans le temps ayant un seuil d'objets en communs suffisamment grand pendant une durée de temps exacte ou minimale selon l'approche choisie. Il existe dans la littérature plusieurs études de clusters d'objets mobiles comme les clusters en mouvement (Kalnis et al., 2005), les Flocks (Li et al., 2011) (Gudmundsson & Van Kreveld 2006) (Benkert et al., 2008), les Convois (Jeung et al., 2008a; Jeung et al., 2008b) et les Swarm (Li et al., 2010a). Les Flocks sont souvent utilisés dans plusieurs applications pour la détection automatique de sous-ensembles d'objets mobiles suivant des trajectoires voisines pendant une durée de temps prédéfinie. L'algorithme Flock se base sur la détection de clusters de rayons fixes ayant un nombre d'objets mobiles minimal pendant des instants de temps successifs t=1, 2,3. Une extension intéressante de l'algorithme Flock est l'algorithme Convoy. En effet, il suppose que les objets se déplaçant ensemble peuvent sortir du groupe ou y entrer à tout moment (Figure 2-21). Par comparaison à l'algorithme Flock, l'algorithme Convoy permet la détection de clusters de formes arbitraires en intégrant la notion de liaison de densité (Cf. section 2.2.1.3). L'algorithme Convoy demande de transformer les trajectoires discrètes en trajectoires continues dans le temps. Les trajectoires traitées par L'algorithme Convoy doivent êtres continues pour pouvoir faire des groupements à n'importe quel temps.

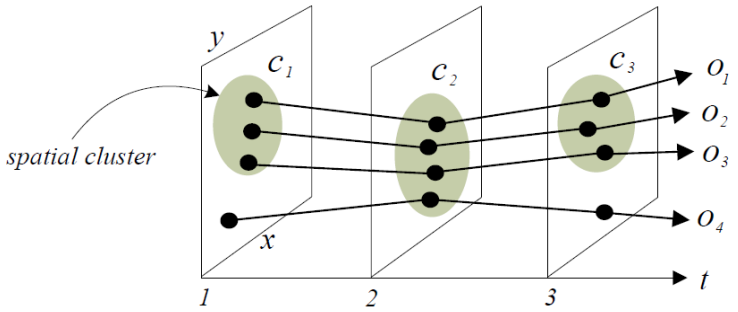


Figure 2-21 : Détection de convois pour un intervalle de temps et un nombre minimal d'objets égal à 3.

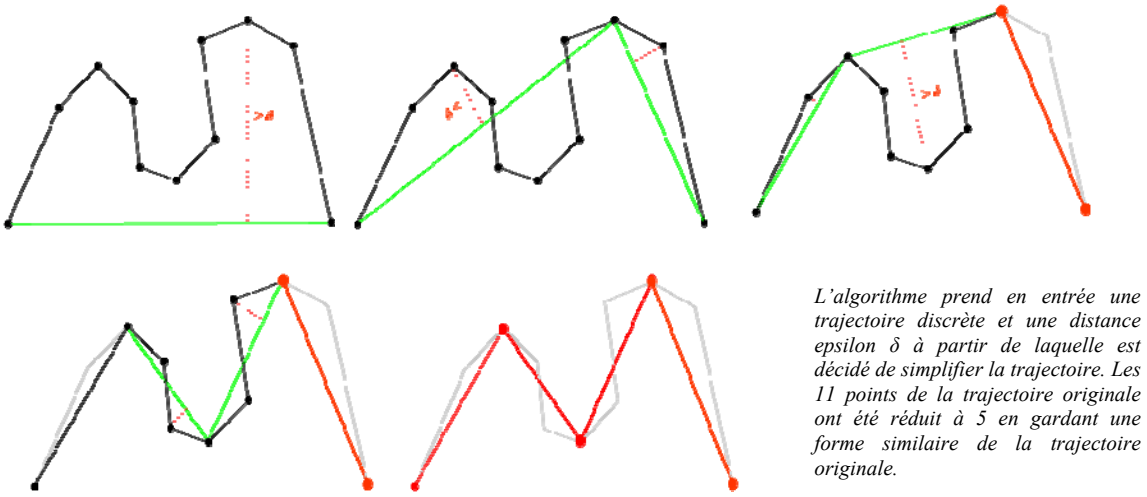
La détection de convois, peut avoir plusieurs applications pratiques comme la détection des troupeaux d'animaux, la détection des manifestations dans les foules et la découverte de navires navigant ensemble.

En s'inspirant des études de détection de clusters d'objets mobiles, plusieurs algorithmes ont été proposés comme les algorithmes de base MC1, MC2 et MC3 développé par Kalnis (Kalnis et al., 2005), BFE (Basic Flock Evaluation) (Vieira et al., 2009) pour la détection de Flock et les algorithme CMC (Coherent Moving Clusters), CuTS (Convoy Discovery using Trajectory Simplification), CuTS+ et CuTS* (Jeung et al., 2008a, Jeung et al., 2008b) pour la découverte de Convoy. L'algorithme CuTS propose une simplification des trajectoires en entrée, les transformant de trajectoires discrètes en trajectoires continues dans le temps. Cette simplification permet aux positions d'êtres définies à tout moment et donc la possibilité de détecter les clusters de mouvement à chaque instant. De plus, cet algorithme offre de meilleures performances que CMC en termes de rapidité de traitement à cause de la simplification des données trajectoires.

Ces algorithmes de détection de convois proposés par l'équipe de H. Jeung différent selon la phase de simplification choisie. CuTS utilise l'algorithme de Douglas Peucker (Douglas and Peucker, 1973) pour la compression de données et la simplification des trajectoires par généralisation. Il permet de ne garder que les positions caractéristiques ou significatives. Formellement, étant donné une trajectoire représentée par la polyligne suivante $o = <p_1, p_2, ..., p_T>$, et un paramètre de tolérance δ donné en entrée, l'objectif de la simplification de trajectoires est de proposer une autre polyligne o' contenant moins de points que la trajectoire originale et s'écartant d'au plus δ de la trajectoire originale o.

Comme illustré sur la Figure 2-22, l'algorithme de Douglas Peucker (DP) relie avec un segment de droite les deux points extrêmes de la trajectoire. Il calcule la distance entre le point le plus éloigné de la trajectoire et ce segment de droite. Si cette distance est supérieure à epsilon δ (seuil de précision), il considère le point le plus éloigné comme le nouveau point extrême de la trajectoire, sinon c'est le segment de droite qui devient la trajectoire simplifiée. La trajectoire est coupée récursivement jusqu'à ce que la condition de distance ne soit plus vérifiée pour toute la trajectoire.

Le fonctionnement de l'algorithme de Douglas-Peucker est illustré à la Figure 2-22.



L'algorithme prend en entrée une trajectoire discrète et une distance epsilon δ à partir de laquelle est décidé de simplifier la trajectoire. Les 11 points de la trajectoire originale ont été réduit à 5 en gardant une forme similaire de la trajectoire originale.

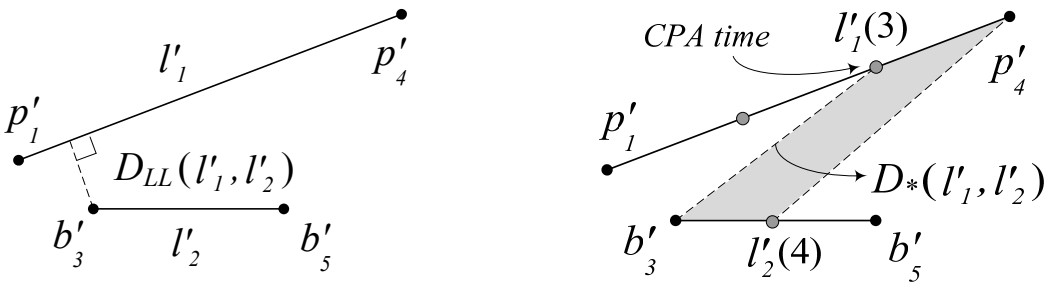
Figure 2-22 : illustration de l'algorithme Douglas-Peucker sur une simplification de trajectoire (source Jeung[61]).

Pour une simplification plus rapide, l'algorithme CuTS+ intègre l'algorithme DP+ qui identifie les points extrêmes dichotomiquement pour arriver plus rapidement à la simplification optimale. Cependant l'algorithme DP original et son extension de simplification ne prennent pas en compte la dimension temporelle dans la simplification, ce qui a incité H. Jeung et son équipe à proposer CuTS*. CuTS* intègre DP* (Meratnia

[61] Présentation de Hoyoung Jeung, Man Lung Yiu, Xiaofang Zhou, Christian S. Jensen, Heng Tao Shen, "Discovery of Convoys in Trajectory Databases", PVLDB, 1(1):1068-1080, 2008

& De By 2004) qui calcule les distances entres les positions de trajectoires et les segments de droite en considérant le temps (Figure 2-23).

Cet algorithme offre une meilleure efficacité selon les études et tests effectués par H. Jeung. La Figure 2-23 montre la différence entre le calcul de distance entre l'algorithme de DP classique et son amélioration DP* intégrant la dimension temps dans la simplification. Le calcul de distance classique utilisé par DP entre deux segments de trajectoires est basé sur la plus petite distance perpendiculaire entre les deux segments originaux et représentatifs. Cette distance DLL (l'1, l'2) est représentée sur la Figure 2-23-(1).



(1) Algorithme DP (2) Algorithme DP*

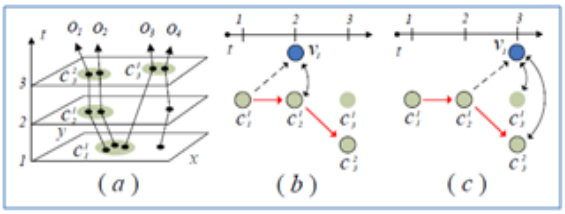
Figure 2-23 : Différence entre le calcul de distance des algorithmes DP et DP*. Dans (1) une distance perpendiculaire est calculée et dans (2) une distance entre 2 positions ayant le même timestamp (Jeung et al., 2008a).

Les algorithmes CMC, CuTS, CuTS+ et CuTS* prennent en entrée, un ensemble de trajectoires de N objets, un seuil de distance e minimum pour le groupement, un nombre minimum d'objets m pour former un convoi et une durée de temps k de déplacement de convoi. L'objectif de ces algorithmes est la découverte rapide et efficace des groupes d'objets se déplaçant ensemble en se basant sur la densité de trajectoires par rapport à la distance e et le nombre de trajectoires m durant k instants consécutifs. Même si ces algorithmes ont le même principe, chacun à ses spécificités.

CMC (Coherent Moving Cluster) est un algorithme basé sur une simple technique de détection de convois. Tout d'abord, l'algorithme utilise une interpolation linéaire des

positions de trajectoires pour compléter celles qui manquent par des positions virtuelles à des instants t donnés. Ensuite, un groupement des objets est effectué à chaque instant t en utilisant la connexion par densité (Cf. 2.2.2 du chapitre 2) pour détecter les clusters possibles. Le rayon utilisé est e. Après avoir détecté ces clusters, l'algorithme cherche les convois candidats qui contiennent des positions communes entre des clusters consécutifs pendant k instants. Un convoi est actuel si au moins les k clusters candidats partagent un nombre suffisant de trajectoires communes. Pour que le nombre soit suffisant, le nombre de trajectoires partagées par les clusters candidats doit être supérieur à m.

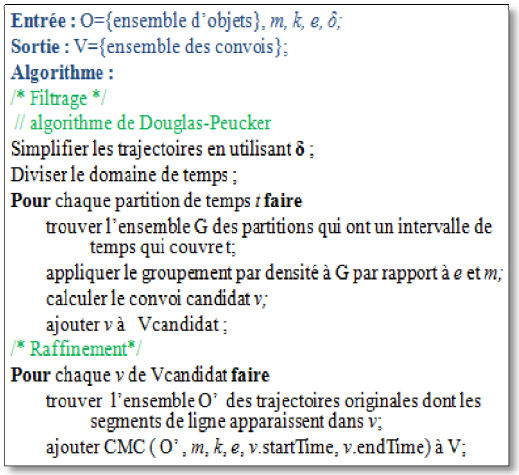
La Figure 2-24 illustre un cas d'exécution de l'algorithme CMC.



Timestamp	Clusters	Candidate set V
t_1	c_1^1	$v_1 = c_1^1$
t_2	c_2^1	$v_1 = c_1^1 \cap c_2^1$
t_3	c_3^1, c_3^2	$v_1 = c_1^1 \cap c_2^1 \cap c_3^2, v_2 = c_3^1$

Figure 2-24 : Exemple illustrant l'exécution de l'algorithme CMC pour m=2 ; k=3 ; e=◄► .

- CuTS (Convoy discovery Using Trajectory Simplification) est une amélioration de l'algorithme CMC pour réduire sa complexité. CuTS est composé de deux phases, une phase de filtrage et une phase de raffinement (Algorithme 2-6). Dans le filtrage, les trajectoires sont simplifiées en utilisant l'algorithme de Douglas Peucker (DP) qui utilise comme paramètre la tolérance δ définie par l'utilisateur. Les partitions de trajectoires trouvées après simplification sont regroupées pour calculer les convois candidats. Dans la phase de raffinement, l'algorithme CMC est repris avec une amélioration. Avec la simplification des trajectoires, il n'y a pas besoin de calculer les interpolations pour compléter les positions de trajectoires manquantes.



Entrée : O={ensemble d'objets}, *m*, *k*, *e*, *δ*;
Sortie : V={ensemble des convois};
Algorithme :
/* Filtrage */
// algorithme de Douglas-Peucker
Simplifier les trajectoires en utilisant **δ** ;
Diviser le domaine de temps ;
Pour chaque partition de temps *t* **faire**
trouver l'ensemble G des partitions qui ont un intervalle de temps qui couvre t;
appliquer le groupement par densité à G par rapport à *e* et *m*;
calculer le convoi candidat *v*;
ajouter *v* à Vcandidat ;
/* Raffinement*/
Pour chaque *v* de Vcandidat **faire**
trouver l'ensemble O' des trajectoires originales dont les segments de ligne apparaissent dans *v*;
ajouter CMC (O' , *m*, *k*, *e*, *v*.startTime, *v*.endTime) à V;

Algorithme 2-6 : Algorithme CuTS.

- CuTS+ améliore la phase de simplification dans CuTS, en utilisant l'Algorithme de Douglas Peucker plus (DP+). DP+ est plus rapide que DP, il diminue l'erreur en cas de groupement car la tolérance actuelle obtenue par DP+ est toujours plus petite que celle obtenue par DP mais un point faible de cet algorithme est qu'il ne conserve pas bien la forme de la trajectoire originale.
- CuTS* est une extension temporelle de CuTS, il introduit la notion du temps lors de la phase de filtrage, en utilisant la version améliorée de Douglas Peucker de base appelé DP* pour la simplification. Cette notion est aussi intégrée lors du groupement en utilisant une fonction de calcul de distance qui respecte la position de l'objet par rapport à l'axe de temps.

2.2.3.1.5. Détection des périodiques

La découverte de mouvements périodiques est intéressante dans l'analyse de comportements des objets mobiles. La périodicité peut être utilisée par exemple pour la prédiction de mouvements habituels, la détection des événements anormaux et pour résumer les longs historiques de mouvements. Parmi les méthodes récentes qui ont été proposées dans la littérature pour la découverte des mouvements périodiques, on trouve Mining Periodic Behaviors for Moving Objects (Li et al., 2010b). Li et son équipe ont proposé cette méthode dans l'objectif de résoudre le problème de détection des comportements périodiques des objets en mouvements. Ce problème est composé de deux sous-problèmes, à savoir la détection des périodes dans les mouvements et la découverte des comportements périodiques.

Les données sur les mouvements observés sont généralement générées à partir de plusieurs comportements périodiques entremêlés, ayant différentes périodes, associés à plusieurs emplacements de référence et ayant plusieurs formats de dates (mois, jours, heures, etc.). Sans oublier le fait que les objets mobiles n'ont pas souvent les mêmes trajectoires ce qui rend difficile la détection des périodicités et leur prédiction. La plupart des méthodes de détection de périodiques appliquent la transformée de Fourier directement sur les séquences de déplacements. Etant donné que les résultats de la transformée de Fourier sont sensibles au bruit[62], la découverte de périodiques peut échouer. Pour résoudre ce problème, la méthode proposée par l'équipe de Li commence par chercher toutes les localisations de références qui sont des zones fréquemment visitées. L'observation des mouvements à partir de ces localisations de référence permet de faciliter la découverte des périodiques.

L'algorithme développé pour supporter cette méthode est appelé Periodica. L'algorithme est en deux étapes : une étape de détection des périodiques à partir des séquences de mouvements données en entrée et une étape de découverte de comportements périodiques (Algorithme 2-7).

Entrée : $Loc=loc_1\ loc_2\ \ldots\ locn$;
Sortie : un ensemble des comportements périodiques;

Algorithme :
/* Etape 1 : détection des périodes */
Trouver les localisations de référence $O=\{o_1, o_2, \ldots, o_d\}$;
Pour chaque localisation de réf. **Faire**
  Transformer le mvt. En une séquence binaire;
  Trouver les périodes dans chaque localisation de réf. *oi;*

/* Etape 2 : Découverte de comportements périodiques */
Pour chaque période *T* **Faire**
  Diviser le mvt. O_T = {localisation de réf. de période *T*};
  Construire la séquence de mvt. symbolisée *S* à partir de O_T;
  Trouver les comportements périodiques dans *S*;

Algorithme 2-7 : Algorithme Periodica.

Dans la première étape, l'algorithme commence par identifier les localisations de référence en utilisant une méthode de calcul de densité par noyau (Kernel Density Estimation[63]). Ces localisations de référence vont faciliter la découverte des périodes. En effet, en observant le mouvement à partir de ces localisations, les périodes sont plus faciles à percevoir. Ensuite pour chaque emplacement de référence identifié, l'algorithme cherche toutes les périodes possibles en appliquant la transformée de Fourier et l'auto-corrélation circulaire sur la transformation binaire de la séquence de mouvements effectuée au préalable.

[62] Veut dire ici des trajectoires non utiles à la découverte de comportements périodiques.
[63] KDE est une méthode de calcul de densité non-paramétrique.

Dans la deuxième étape de l'algorithme, pour chaque période T trouvée, l'algorithme segmente le mouvement par ces périodes. Ensuite, il considère seulement les localisations de référence et construit une séquence symbolisée de mouvements à partir de ces localisations. C'est à partir de cette séquence symbolisée que Periodica cherche à découvrir les comportements périodiques en utilisant une méthode de groupement (voir 2.2.1.3). Les clusters identifiés à la fin représentent les comportements périodiques.

Pour bien comprendre l'algorithme Periodica, prenons un exemple simple de découverte de périodicités dans un jeu de données de déplacement d'une abeille (Figure 2-25) présenté par Z. Li[64]. L'utilisation de Periodica sur ce jeu de données va permettre en premier lieu d'identifier les localisations de référence, à savoir dans les ruches et en dehors des ruches, puis de transformer les séquences de mouvements en des séquences binaires par rapport aux zones de référence (voir Figure 2-26).

L'utilisation de la Transformée de Fourier sur les séquences binaires de mouvements permet de détecter la présence de périodiques. Pour avoir les valeurs de ces périodiques, l'Autocorrélation Circulaire est utilisée.

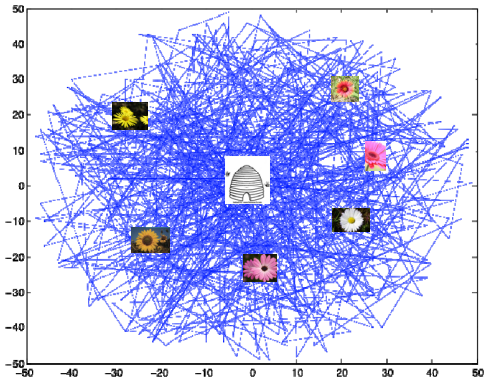


Figure 2-25: mouvement d'une abeille (Li et al., 2010).

[64] http://faculty.ist.psu.edu/jessieli/Publications/KDD10-ZLi-PeriodMovObj.ppsx

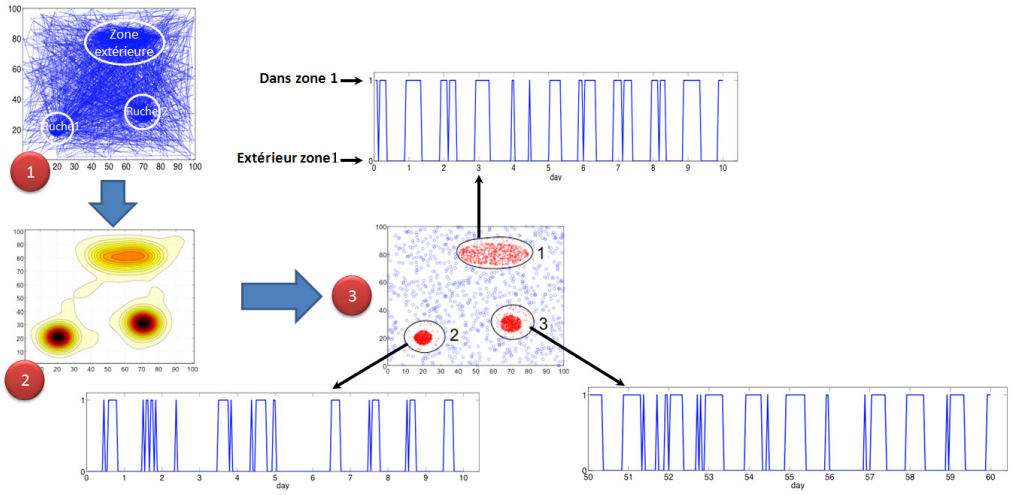


Figure 2-26 : Première étape de Periodica - (1) détection des localisations de références, (2) transformation des mouvements en une séquence binaire, (3) trouver les périodes (Li et al., 2010).

Après avoir identifié la période du comportement de l'abeille qui est de 24h, les mouvements sont symbolisés par les localisations de références, segmentés en périodes de 24h et regroupés par une méthode de classification hiérarchique ascendante (Figure 2-27). Initialement, chaque segment est un comportement et la distance entre les comportements est calculée en utilisant la mesure de dissimilarité divergente de Kullback-Leibler. Le comportement périodique est modélisé sous forme d'une matrice de probabilités de présence dans les zones de références (appelées spot sur la Figure 2-27).

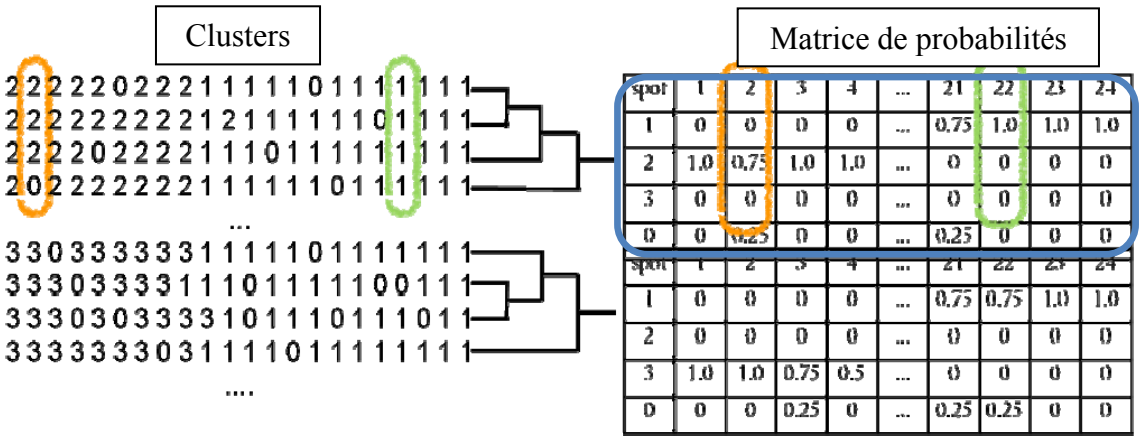


Figure 2-27 : Deuxième étape de Periodica – Clustering des segments de mouvements, construction de la matrice de probabilités et découverte des comportements périodiques (Li et al., 2010).

La découverte du nombre de comportements périodiques est effectuée par l'utilisation de la mesure de pureté du groupement qui est l'erreur de représentation des segments dans un cluster. Deux comportements périodiques ont été identifiés, le premier est un déplacement entre la ruche 1 et la zone extérieure et le deuxième entre la ruche 2 et la zone extérieure. Dans les deux comportements, l'abeille passe 8h dans la ruche et 16h à voler à proximité de la ruche pour chercher de la nourriture.

Un autre exemple d'application intéressante est celui de la découverte d'un comportement migratoire d'un aigle chauve en Amérique du Nord. Les données sur lesquelles l'équipe de J. Han ont fait le test représente le déplacement d'un aigle pendant une durée de 3 ans. La Figure 2-28 montre le résultat de Periodica sur ce jeu de données. La partie (a) de la figure montre les localisations de l'aigle sur Google Earth, la partie (b) illustre les trois localisations de références qui sont détectées par analyse de densité et la partie (c) décrit le comportement périodique de l'aigle. Selon ce résultat, l'aigle chauve reste à la localisation 1 (spot 1) de décembre à mars, il commence à se déplacer vers la localisation 2 (spot 2) en mars, il reste dans cette localisation jusqu'à la fin du mois de mai, après il migre vers la localisation 3 (spot 3) en été. Ensuite, à partir de septembre il revient vers la localisation 2 et y reste de mi-octobre à mi-novembre puis revient à la localisation 1 et ainsi de suite.

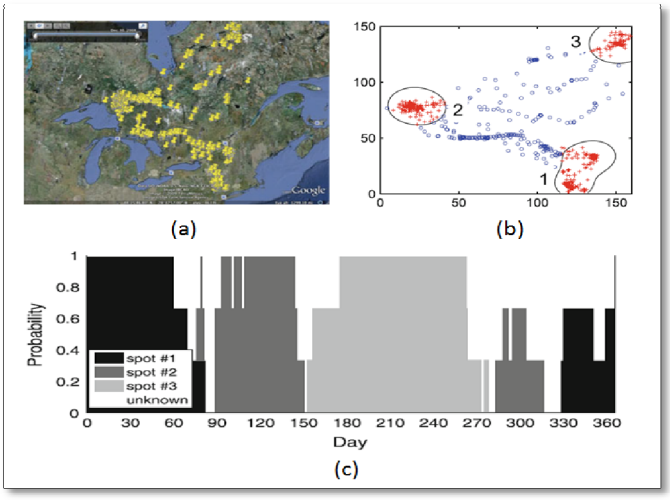


Figure 2-28 : Découverte de comportements périodiques dans le déplacement d'un aigle chauve. (a) Les données affichées sur Google Earth. (b) Découverte des localisations de référence. (c) Le comportements périodiques de l'aigle chauve (Li et al., 2010).

2.2.3.2. Espace d'évolution contraint par un réseau

D'autres méthodes de fouille de données d'objets mobiles, comme le clustering de trajectoires, sous-trajectoires (Kharrat et al., 2008a; Kharrat et al., 2009) (Lai et al., 2007) (Meng, 2007) (Chen et al., 2007) et de découverte de patrons de mobilités (Kharrat et al., 2013) sont utilisées pour les objets mobiles dont le mouvement est soumis à des contraintes de trajectoires (réseau routier, réseau ferré, etc.). Ces objets ne peuvent pas se déplacer librement dans l'espace mais doivent suivre un réseau qui influence la forme et la vitesse de leur mouvement (Kharrat et al., 2013). On parle dans ce cas, de trajectoires contraintes par un réseau (Kharrat et al., 2008b). Pour exploiter ces données de mouvements, plusieurs approches de représentations des données d'objets mobiles sont proposées. Parmi ces approches, nous pouvons citer les deux plus communément utilisées. La première se base sur un découpage de l'espace de déplacement en une grille avec des cases de tailles fixes. La deuxième se base sur une représentation sous forme de graphes représentant le réseau sur lequel se déplacent les objets. Cette dernière représentation est plus intéressante car elle permet d'avoir des résultats précis et complets reflétant la réalité du terrain (Lai et al., 2007). Dans la représentation sous forme de graphe, les arrêtes du graphe sont des segments du réseau et les nœuds sont des interconnections entre les segments. Ces segments peuvent être orientés ou non.

Plusieurs applications ont besoin de ce type de méthodes de fouille de données comme la gestion du trafic routier, la détection des congestions du trafic routier et la découverte de bouchons.

Dans notre problématique, les deux approches de fouille de données peuvent être appliquées. En effet, il est possible de considérer les navires comme des navires soumis à des contraintes de routes, des objets se déplaçant dans un espace ouvert ou comme une combinaison des deux. La combinaison peut être par rapport à des parties de l'espace comme par exemple être soumis à des contraintes de routes dans les approches d'entrée/sortie d'un port et considérer que le déplacement est ouvert en pleine mer. Il est possible aussi d'imaginer une combinaison par rapport au type du navire. Les Tankers ont tendance à suivre des routes bien définies alors que les navires de pêches par exemple ont des déplacements spécifiques qui dépendent de plusieurs critères (type de pêche, zones de pêches, météorologie, etc.).

Dans la suite de ce travail, nous avons fait le postulat que le déplacement des navires se fait plutôt dans un espace ouvert non soumis à des contraintes de routes pour plusieurs raisons :

- Les routes maritimes ne sont pas matérialisées. Dans les méthodes destinées aux objets contraints par un réseau, il faut au préalable de l'analyse, construire le réseau de routes,
- Des types de navires se déplacent d'une manière spécifique sans suivre des routes particulières comme les navires de pêche et de plaisance.

2.2.4. La fouille de données du trafic de mobiles

Pour pouvoir distinguer ces deux derniers domaines, à savoir la fouille de données d'objets mobiles et la fouille de données du trafic, il faut observer l'objet sur lequel porte l'analyse. Dans la fouille de données d'objets mobiles, l'analyse porte sur le mobile et ses déplacements. Dans la fouille de données du trafic, l'objectif est d'analyser les flux d'objets mobiles qui passent par des segments de réseaux (routier, ferroviaire, etc.). La différence peut être cernée aussi dans les données de capteurs, étant donné que la fouille de données d'objets mobiles utilise des données individuelles issues de capteurs embarqués alors que la fouille de données du trafic, utilise des données du trafic mesurées en différentes parties du réseau.

On enregistre une abondance de données sur le trafic routier, maritime ou aérien. Google Maps par exemple fournit des informations en temps-réel sur la circulation routière comme on le voit sur la Figure 2-29. Il est même possible de faire des prédictions de circulation par jour et par heure en se basant sur les observations passées. Ces données récoltées peuvent êtres exploitées dans plusieurs applications de fouille de données du trafic comme les systèmes de surveillance du trafic (Lu et al., 2006), la découverte des événements de congestion, la planification des routes, la découverte des routes rapides (Awasthi et al., 2005), populaires, etc. L'objectif de ce domaine de fouille de données est de résumer comment

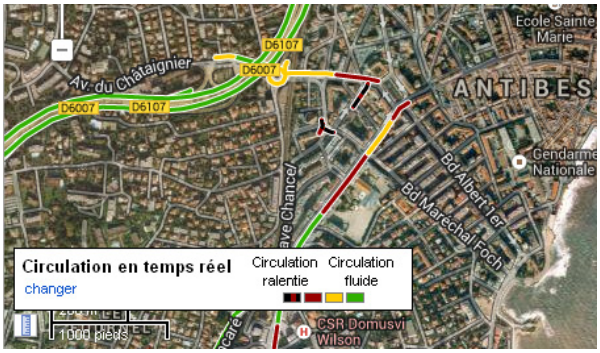


Figure 2-29 : Affichage sur Google Maps du Trafic routier à Antibes en temps-réel.

les objets mobiles se déplacent dans différents intervalles de temps (journée, semaine, etc.), quels sont leurs comportements typiques dans des régions et à des moments spécifiques, etc.

Nous allons exposer dans les sous-sections suivantes, deux méthodes de fouille de données utilisées pour l'exploration de données du trafic : les motifs de trajectoires (T-patterns) et la détection des congestions (T-Flock). Les trajectoires sur lesquelles porte l'exploration sont des séquences de localisations d'objets mobiles et de temps de transition d'un certain point du réseau à un autre.

2.2.4.1. Les motifs de trajectoires

Les motifs de trajectoires sont des séquences de zones avec des transitions étiquetées dans le temps : $T = Z_0 \rightarrow_{\alpha 0} Z_1 \rightarrow_{\alpha 1} .. \rightarrow_{\alpha 2} Z_n$, tel que $Z = \{ Z_{0,..,} Z_n\}$ est l'ensemble des zones et $\alpha = \{\alpha_{1,...,} \alpha_n\}$ est l'ensemble des temps de transition entre les zones. Ces motifs de trajectoires sont utilisés pour résumer le comportement du trafic, découvrir les régions de congestions et la prédiction des déplacements futurs, etc. Trajectory Pattern Mining (T-Patterns) proposé par l'équipe *Knowledge Discovery and Data Mining Laboratory (KDD Lab)* de l'Université de Pise (Giannotti et al., 2007), est un algorithme qui permet l'extraction de ce type de motif. Il prend en entrée des données D (latitude, longitude, timestamp), le support de trajectoires, le support de temps et le rayon de voisinage et il retourne un T-Patterns. Le support de trajectoires est le pourcentage de trajectoires qui doivent composer le T-Patterns, le support de temps est le temps minimum d'une transition et le rayon de voisinage $N(X_i,Y_i)$ est le disque englobant les zones denses. Pour comprendre comment l'algorithme fonctionne, un exemple d'extraction d'un motif T-Pattern est représenté sur la Figure 2-30.

L'algorithme suit trois étapes pour extraire les motifs de trajectoires :

- Calculer les RoI (Region of Interest) qui sont les zones denses,
- Convertir les données (x, y, t) en (RoI, t),
- Trouver les séquences de transitions satisfaisant le support de trajectoires et de temps.

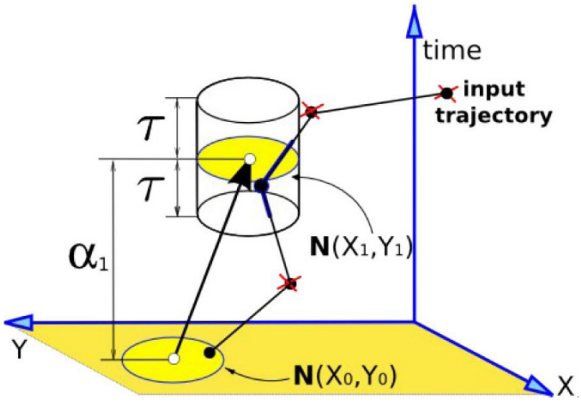


Figure 2-30 : Illustration de la méthode d'extraction d'un motif de trajectoire à partir d'une trajectoire (Giannotti et al., 2007)[65].

2.2.4.2. La détection des congestions

La détection de congestions ou des embouteillages à partir des données de trafic peut être vu comme un problème de détection de groupes d'objets se déplaçant ensemble et lentement. L'équipe de Giannotti (Ong et al., 2011) ont proposé une méthode de détection des embouteillages basée sur une extraction des Flocks (Cf. section 2.2.3.1.4). Cette méthode enrichit le concept de Flock par un support de vitesses pour pouvoir identifier les objets qui se déplacent à proximité et ayant des vitesses inférieures au support (T-Flock).

2.3. Prototypes d'analyse de comportements

Certaines équipes de recherche qui travaillent activement dans le domaine de la fouille de données, comme l'équipe de Jiawei Han du *Department of Computer Science University of Illinois* et l'équipe de Fosca Giannotti du laboratoire *Knowledge Discovery and Data Mining* de l'université de Pise ont développé des prototypes regroupant les méthodes de fouille de données d'objets mobiles développées en interne. Ces prototypes servent de preuve de concept et de vitrine aux travaux réalisés. Nous décrivons ci-après, les deux prototypes MoveMine et M-Atlas qui ont été développés par ces deux équipes.

[65] Trajectory Pattern Mining, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.187.6467&rep=rep1&type=pdf

2.3.1. MoveMine

Le prototype MoveMine (Li et al., 2010c) a été développé dans le cadre d'un projet avec Boeing pour une éventuelle utilisation dans le transport aérien. Un démonstrateur du prototype avait été mise en ligne[66] sans la possibilité de téléchargement du code source ou de son exécutable. Les éditeurs de MoveMine sont tenus à une obligation de confidentialité du projet.

L'architecture système de MoveMine, se compose de trois interfaces (Figure 2-31) : une interface de collection et nettoyage des données, une interface de fouille de données et une interface de visualisation des données sources et des résultats.

Dans l'interface Collection and Cleaning, un ensemble de données de déplacements d'animaux, de véhicules et des événements climatiques sont collectés et prétraités pour permettre leur exploration par les algorithmes de fouille de données. Les données en entrée peuvent contenir des bruits, des imprécisions, des incohérences : c'est pour cela qu'un module de nettoyage est ajouté dans l'architecture du prototype.

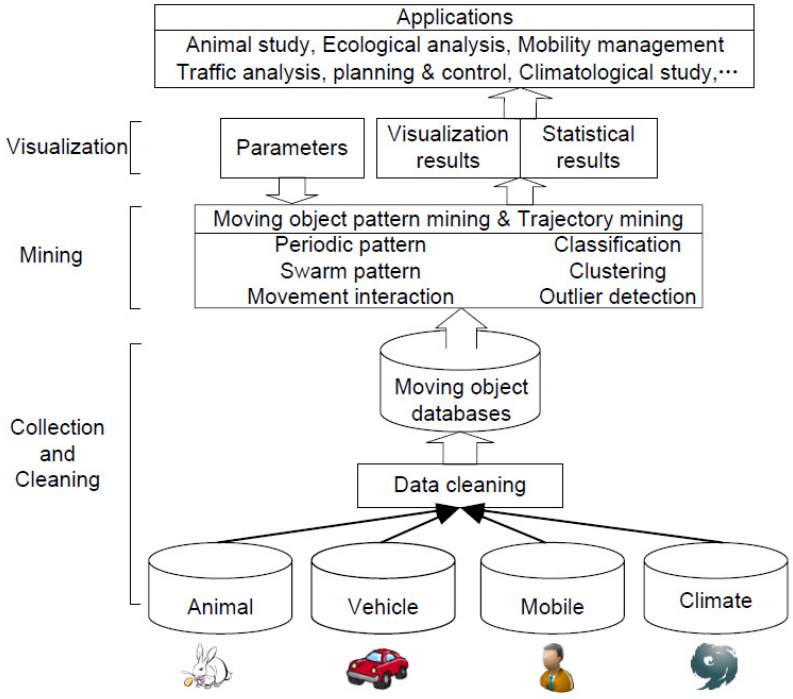


Figure 2-31 : Architecture du prototype MoveMine (Li et al., 2011).

[66] http://dm.cs.uiuc.edu/movemine/

Dans l'interface Mining, un ensemble d'algorithmes de fouille de données d'objets mobiles développés récemment par l'équipe de J. Han ont été intégrés. Un résumé des fonctionnalités de ces algorithmes est représenté sur la Figure 2-32.

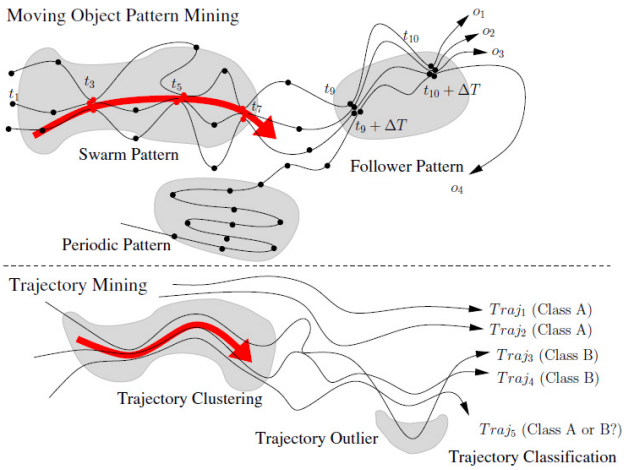


Figure 2-32 : Les fonctionnalités principales de MoveMine

(Li et al., 2011).

A partir de l'interface Visualization, les utilisateurs du système peuvent entrer les paramètres, lancer l'exécution d'un algorithme, afficher les données sur Google Maps, Google Earth et interagir avec les résultats.

Le démonstrateur mis sur internet montre son utilisation pour plusieurs domaines d'application comme l'étude des déplacements d'animaux, l'analyse écologique, la gestion de la mobilité, l'analyse du trafic, l'étude climatique, etc. Il permet de révéler des clusters de trajectoires, de détecter des trajectoires aberrantes (outlier trajectory), d'identifier des comportements périodiques et des mouvements collectifs (Figure 2-33). Le prototype MoveMine permet d'évaluer les algorithmes proposés par l'équipe.

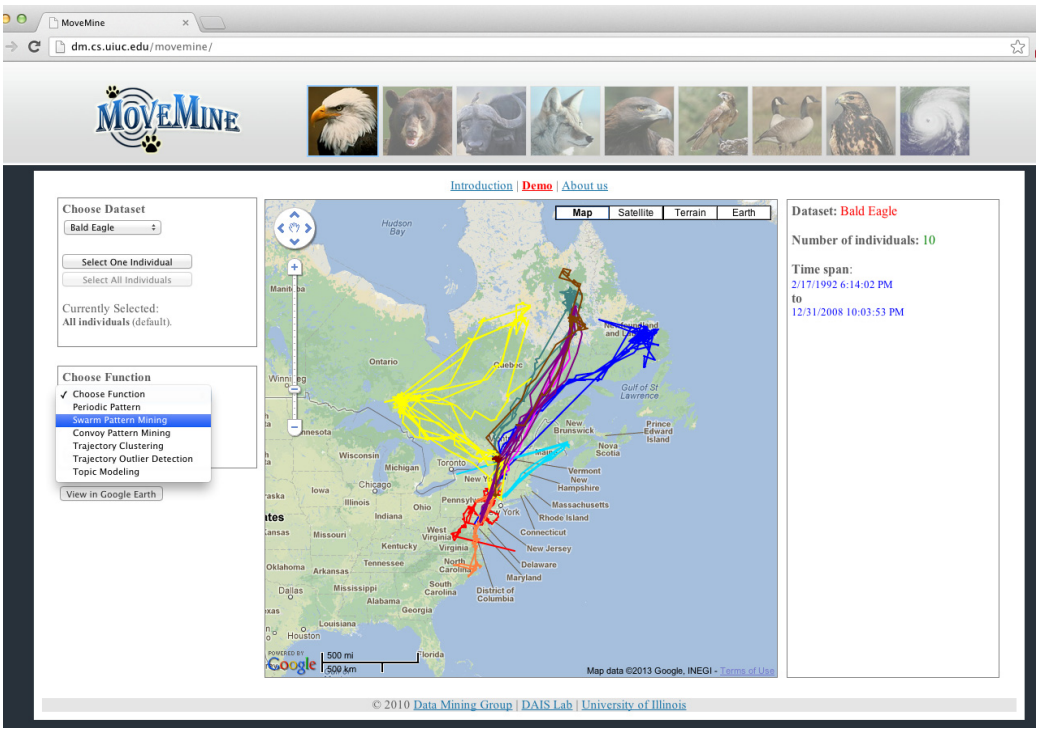


Figure 2-33 : Démonstrateur MoveMine-Exemple de détection de mouvements collectifs à partir d'un jeu de données de déplacement d'aigles (Li et al., 2010c).

2.3.2. M-Atlas

M-Atlas est un prototype qui a été développé et mis en téléchargement libre[67] par l'équipe de Giannotti (Giannotti et al., 2011). Il est centré sur le concept de trajectoire sur laquelle porte les différentes analyses. Ce prototype se propose d'analyser le trafic et la mobilité des objets par l'interrogation et l'exploration des données de trajectoires construites à partir des traces GPS brutes. Il intègre un mécanisme de transformation de données, de construction de trajectoires, de stockage de trajectoires/motifs/modèles, d'interrogation de trajectoires/motifs/modèles, intègre des outils de fouille de trajectoires pour l'extraction de motifs, de construction de modèles et une interface de visualisation. Les motifs extraits et les modèles construits sont combinés afin d'améliorer la découverte de connaissances de mobilité. La découverte de ces connaissances est une interaction entre deux mondes, à savoir celui des données (trajectoires) et celui des modèles et des

[67] http://www.M-Atlas.eu/

motifs de mobilité[68] (Giannotti et al., 2011). Il y a quatre motifs de mobilité et trois modèles qui sont supportés par M-Atlas.

Les motifs représentés sur la Figure 2-34 décrivent dans l'ordre, le clustering de trajectoires (T-Cluster), les T-Patterns, les T-Flock (2.2.4.2) et le flux d'objets qui se déplacent d'une région à une autre (T-Flow).

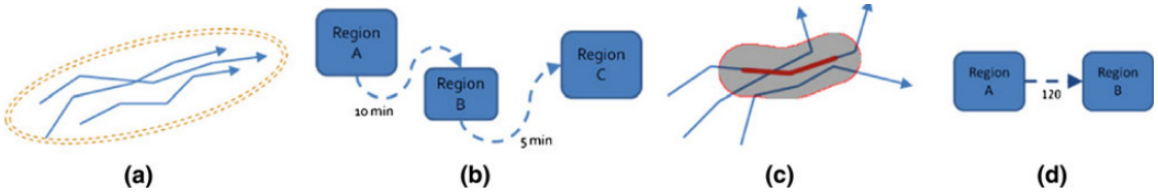


Figure 2-34 : Les motifs de mobilité supportés par M-Atlas – (a) T-Cluster, (b) T-Pattern, (c) T-Flock, (d) T-Flow (Giannotti et al., 2011).

Les modèles représentés sur la Figure 2-35 quant à eux, décrivent respectivement, l'accessibilité des partitions de trajectoires selon un seuil de distance (Accessibilité), un ensemble de motifs T-Patterns (T-PTree) et un graphe de transition labélisé par le nombre de passages entre des régions origine et destination (T-O/DMatrix).

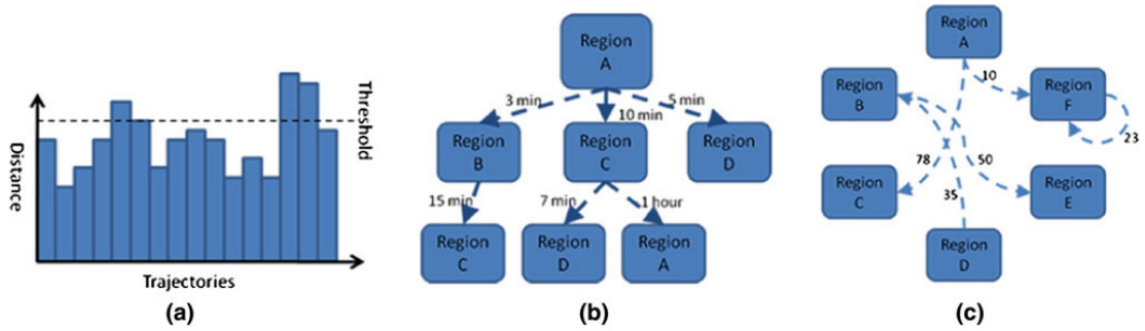


Figure 2-35 : Les modèles supportés par M-Atlas – (a) Accessibilité, (b) T-PTree, (c) T-O/DMatrix (Giannotti et al., 2011).

[68] Comportement commun d'un groupe ou un sous-groupe de trajectoires qui a été obtenu par fouille de données (Giannotti et al. 2011).

Le prototype a été développé dans l'objectif de répondre à un certain nombre de questions sur la mobilité des objets, du genre : Quels sont les motifs fréquents de déplacement ? Comment les événements influencent-t-il la mobilité ? Comment prévoir les bouchons de circulation ? Comment découvrir les embouteillages et les congestions ? etc. Il est à remarquer que les questions auxquelles répond M-Atlas sont plutôt du domaine de la fouille de données du trafic. Un exemple d'extraction de T-Patterns et du motif T-O/DMatrix est représenté sur la Figure 2-36.

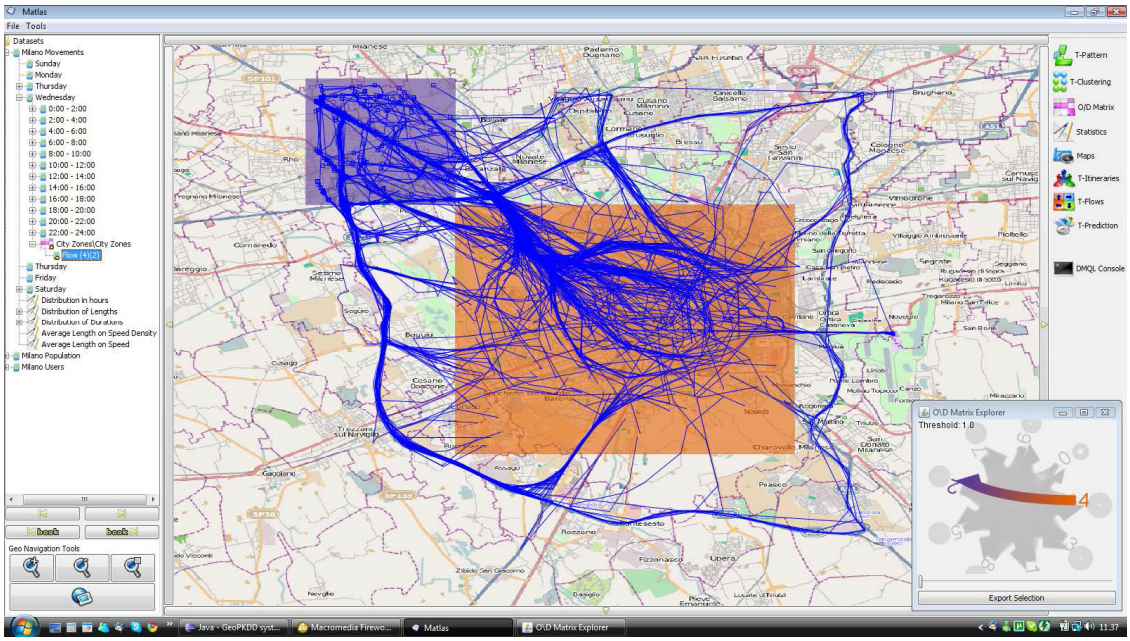


Figure 2-36 : Interface M-Atlas du résultat de T-Pattern entre le centre-ville (couleur orange) et le nord-ouest. La petite fenêtre à droite est la matrice de transition T-O/DMatrix (Giannotti et al., 2011).

2.4. Synthèse des méthodes de fouille de données

Nous avons exposé dans les sections précédentes un état de l'art sur des problèmes, des méthodes et algorithmes de fouille de données. Dans l'objectif de synthétiser cet état de l'art, nous présentons ci-après un tableau récapitulatif de ces méthodes et algorithmes de fouille de données :

Fouille de données classique			
Problèmes/Méthodes	**Algorithmes**	**Avantages**	**Limites**
Extraction de règles d'association	Apriori	- Réduction de l'espace de recherche en exploitant une structure de données en treillis de Galois, - Exploitation de l'antimonotonie du support pour élaguer le treillis, - Amélioration de la sélectivité par d'autres mesures complémentaires au support et à la confiance, - Simplicité de l'algorithme et efficacité pour une base de données pas très volumineuse.	- Scanne (n+1) fois la base de données /n est la taille du plus long itemset fréquent, - Calcul coûteux de l'heuristique quand les itemsets fréquents sont nombreux.
	TreeProjection	- Utilisation d'un arbre lexicographique pour faciliter la gestion et le comptage des candidats, - Limitation du comptage du support.	- Algorithme peu performant pour de grands volumes de données.
	FP-growth	- Nouvelle structure de données « FP-Tree » proposant un accès plus efficace, - Scanne la base de données 2 fois.	- Nombre important de règles générées.
Arbre de décision	ID3	- Algorithme de référence et l'un des plus préféré en apprentissage automatique, - Non restreint aux attributs binaires.	- Ne supporte pas les variables continues et les valeurs manquantes.
	CART	- Précision et exhaustivité, - Inférence des arbres binaires, - Peut être étendu au traitement des variables continues.	- Non détection de combinaisons de variables, - Besoin d'un échantillon de données de grande taille.
	C4.5	- Prise en compte de variables continues, - Successeur d'ID3.	- Non détection de combinaisons de variables, - Introduction de variables non significatives (induisant un gain nul) dans la construction de l'arbre.

Fouille de données spatiales			
Problèmes/Méthodes	**Algorithmes**	**Avantages**	**Limites**
Clustering de positions / clustering par densité : - Détecter automatiquement le nombre de clusters, - Découvrir les clusters ayant des formes arbitraires.	DBSCAN	- Faible complexité O(nLogn), - Résultats visuels et textuels, - Validation visuelle des clusters générés.	- Rayon de voisinage fixé au préalable, - Résultats pouvant variés d'une exécution à une autre à cause du choix arbitraire de la première position.
	OPTICS	- Unicité du résultat (le même quelque soit l'exécution), - Variation du rayon de voisinage selon la densité de chaque cluster.	- Résultats visuels et non textuels, - Besoin d'un index pour avoir une complexité similaire à celle de DBSCAN.
Fouille de données d'objets mobiles			
Problèmes/Méthodes	**Algorithmes**	**Avantages**	**Limites**
Détection de trajectoires aberrantes	TRAOD	- Complexité en $O(n^2)$/ n : le nombre de partitions de trajectoires, - Découverte de partitions et de trajectoires aberrantes en se basant sur une mesure de distance et de densité, - Simplification des trajectoires sans besoin de paramètres, - Ne requiert pas une phase d'apprentissage automatique.	- Ne considère pas la dimension temps, - Non adapté à des applications dont les objets mobiles sont contraints par un réseau.
Clustering de trajectoires	TRACLUS	- Précis et efficace, - Basé sur une adaptation de l'algorithme DBSCAN.	- Sensibilité des trajectoires représentatives aux décalages des départs et arrivées des partitions de trajectoires, - Partitionnement des trajectoires basé sur un algorithme approximatif (algorithme glouton).
Clustering d'objets mobiles	Flock	- Applicable directement sur des trajectoires discrètes, - Prend en compte la dimension temps.	- La durée de temps des déplacements proches doit être prédéfinie d'une manière exacte, - Clusters de rayon fixe.

	Convoy	- Temps de déplacements proches prédéfini comme un seuil minimal, - Détection de clusters de formes arbitraires, - Prise en compte la dimension temps, - Simplification des trajectoires par une amélioration de l'algorithme Douglas-Peucker intégrant la dimension temps, - Rapide et efficace,	- Besoin de transformer les trajectoires discrètes en trajectoires continues, - N'intègre pas d'autres critères dans le regroupement comme la vitesse par exemple.
Classement de trajectoires	TRACLASS	- Précision de la classification à cause du regroupement basé sur les régions et les trajectoires, - Utilisation des partitions de trajectoires dans la classification, - Application à plusieurs domaines.	- N'intègre pas la dimension temps, - Ne supporte pas encore la classification basée sur des mesures numériques.
Détection des périodiques	Mining Periodic Behaviors for Moving Objects (Periodica)	- Découverte facilitée des périodiques à cause de l'observation des déplacements à partir de zones de référence, - Supporte l'incomplétude des données en entrée, - Résultat non sensible au bruit (trajectoires aberrantes).	- Symbolisation des déplacements avant d'appliquer la transformée de Fourier.

Table 2-2 : Synthèse de méthodes et algorithmes de fouille de données.

2.5. Conclusion

Dans ce chapitre nous avons présenté un panorama des domaines de la fouille de données qui nous intéressent essentiellement pour la suite de ce travail, à savoir, la fouille de données classique, la fouille de données spatiales et la fouille de données d'objets mobiles. Nous avons aussi posé les concepts de fouille de données (algorithmes de fouille de données, paramétrages, etc.) nécessaires à la bonne compréhension de la suite de ce travail. Pour chaque domaine de fouille de données présenté, nous avons détaillé quelques problèmes, méthodes, algorithmes et présenté des exemples d'application quand

cela était nécessaire. Nous avons également présenté deux prototypes de fouille de données à partir desquels nous nous sommes inspirés pour la conception et le développement d'un environnement d'aide à l'analyse de comportements de navires. Enfin, un tableau récapitulatif des méthodes et algorithmes de fouille de données a été présenté à la fin de ce chapitre.

Les prototypes de fouille de données présentés sont soit confidentiels, soit non adaptés pour une utilisation directe ou certains algorithmes implémentés ne répondent à aucune fonctionnalité intéressante pour notre problématique. Dans ce contexte, nous avons été amenés à concevoir et développer un atelier intégrant les méthodes de fouille de données adaptées au contexte maritime et en particulier à l'analyse des comportements à risques de navires en mer.

Dans le chapitre suivant, nous présentons la conception et le développement d'un environnement d'aide à l'analyse de comportements à risques de navires. Nous avons implémenté dans cet environnement un ensemble d'algorithmes de fouille de données adaptés pour l'extraction de connaissances sur les comportements anormaux de navires. Ces comportements anormaux peuvent être interprétés comme potentiellement à risques.

Chapitre 3 : ShipMine : un atelier d'extraction de connaissances pour l'analyse de comportements

3.1. Introduction

Nous souhaitons à partir de ce travail de thèse répondre à la problématique d'aide à l'analyse des comportements à risques de navires en proposant une méthodologie d'extraction de connaissances sur les comportements, basée sur la fouille de données. Pour cela, un ciblage des problèmes, méthodes et algorithmes permettant d'extraire ces connaissances est fondamental. Ensuite, pour valider la méthodologie, les algorithmes ciblés pour l'analyse des comportements de navires vont être testés sur des données réelles de mouvements et d'événements historiques de navires. Ces algorithmes ont été intégrés dans un seul environnement qui va servir de preuve de concept pour montrer l'intérêt de la fouille de données au domaine de la modélisation de comportements potentiellement à risques. Cet environnement reprend les différentes phases de la méthodologie proposée dans ce travail de recherche.

Ce chapitre est organisé en deux parties : la première partie concerne la conception de l'environnement et la seconde, la présentation de son fonctionnement. Dans la partie conception, les utilisateurs potentiels ainsi qu'une analyse de besoins d'extraction de connaissances sur les comportements à risques sont présentés, la sélection des algorithmes et leur intégration dans l'environnement sont détaillées, les données à utiliser dans l'exploration sont aussi présentées et enfin l'architecture du prototype est exposée avec les technologies qui ont été choisies pour son implémentation. Dans la seconde partie de ce chapitre, la présentation du prototype et son fonctionnement sont exposés.

3.2. Conception et réalisation de l'atelier

L'environnement que nous voulons développer est un atelier d'aide à l'extraction de connaissances sur les comportements normaux et anormaux de navires. L'atelier, que nous avons nommé ShipMine, va intégrer un ensemble d'algorithmes et de programmes informatiques visant à automatiser l'extraction de comportements. L'interprétation de ces comportements, va aider à l'acquisition de connaissances sur les risques et leur modélisation.

ShipMine va utiliser la fouille de données pour explorer des données maritimes de déplacement de navires et d'événements historiques. Aujourd'hui, il y a énormément de problèmes, méthodes et algorithmes de fouille de données qui ont été conçus pour

l'extraction de connaissances à partir de données. Les méthodes les plus récentes permettent d'extraire des motifs de mouvements inhabituels, fréquents, périodiques et des résumés à partir de données de déplacements. Dans ce travail, nous nous intéressons à la découverte de connaissances sur les comportements à risques de navires. La difficulté à laquelle nous sommes confrontés est que les méthodes de fouille de données actuelles n'ont pas été conçues pour extraire ce genre de comportements.

Nous rappelons que le comportement étant une combinaison de mouvements dans des situations (Cf. section 2.1 du chapitre 2), il devient alors possible d'initier l'analyse des comportements à partir de motifs de mouvements et de situations pouvant décrire des comportements à risques. L'exploration des données maritimes en utilisant la fouille de données va donc permettre de découvrir des connaissances sur les comportements potentiellement à risques, sous forme de mouvements et de situations.

Pour supporter tout le cycle de vie du développement de notre atelier d'extraction de connaissances basé sur la fouille de données, un modèle en spirale[69] a été retenu. Le choix de ce modèle peut être justifié par le fait qu'il est incrémental, ce qui permet d'avoir des résultats intermédiaires. Ce modèle est bien adapté à notre cas car les résultats ne sont pas connus à l'avance. Les tests des résultats intermédiaires vont enrichir l'atelier au fur et à mesure de l'avancement des développements.

Les phases de développement sont représentées par un schéma en spirales où la première boucle est la définition des besoins, la dernière est la livraison, et entre les deux, un nombre variable de boucles (itérations) qui sont décomposées en quatre phases :

- Détermination des objectifs : choix d'une fonctionnalité, ciblage des algorithmes, acquisition et traitement des données,
- Identification et gestion des risques : données (non accessibles, non exploitables, etc.), algorithme (non adapté, très complexe à implémenter, performances),
- Développements, tests et validation,
- Planification de la prochaine phase : choix d'une nouvelle fonctionnalité.

Dans la suite, la phase de *Détermination des objectifs* et la phase de *Gestion des risques* sont incluses dans les sections *Ciblage et adaptation des algorithmes* (Cf. section

[69] Boehm, 1988, http://weblog.erenkrantz.com/~jerenk/phase-ii/Boe88.pdf

3.2.4) et *Construction de l'espace de données à explorer* (Cf. section 3.2.5). Le développement de l'atelier est traité quant à lui dans la section *Choix technologiques* ci-dessous et le *test et la validation* dans le chapitre suivant.

3.2.1. A qui s'adresse cet atelier ?

Sur l'échelle temporelle, l'atelier permet une analyse *a posteriori* (Cf. section 2 de l'introduction). Les connaissances extraites automatiquement à partir des comportements observés vont aider à modéliser des comportements à risques. Donc pour bien utiliser l'atelier, il faut avoir des compétences en analyse de données et des connaissances dans le domaine maritime.

L'atelier est destiné à des analystes qui seront en charge d'analyser, de raisonner sur les résultats obtenus par l'atelier et de modéliser des comportements à risque à partir de ces résultats. Les autorités maritimes ne sauvegardent pas les déplacements de navires et travaillent juste sur les dernières positions connues. Seuls quelques centres de recherche sauvegardent les historiques de données maritimes dans l'objectif de les utiliser en recherche. Il est difficile donc d'imaginer quels acteurs maritimes actuels vont utiliser cet atelier. Nous supposons que les analystes dont on parle, peuvent êtres des experts maritimes qui ont reçu une formation sur l'utilisation de l'atelier, des scientifiques qui s'intéressent à l'analyse des risques maritimes ou des analystes recrutés spécialement pour utiliser l'atelier. Nous appelons « expert maritime » toute personne capable de valider les résultats de l'atelier, de décider seul ou après approbation de l'intérêt des informations générées et de pouvoir construire des connaissances aidant à modéliser les comportements à risques.

3.2.2. Analyse de besoins

Dans l'objectif de tester et valider notre méthodologie d'extraction de connaissances sur les comportements potentiellement à risques, un ensemble de ces comportements doit être établi pour essayer de cibler des méthodes de fouille de données permettant de les extraire. Il est à rappeler que l'analyse de ces besoins est issue de la littérature.

Pour cela, nous avons choisi deux types de situations et trois types de mouvements pouvant décrire des risques (Figure 3-1). Les deux types de situations (en bleu) sont les zones à risques et les facteurs de risques et les trois types de mouvements (en rouge) sont

les navigations proches, les trajectoires et les routes de navigation (trajectoires types) à risques. Il est possible à partir de ces types de situations et mouvements à risques de générer plusieurs comportements à risques. Les mouvements et les situations qui composent ces comportements vont être définis comme des fonctionnalités que doit assurer notre atelier.

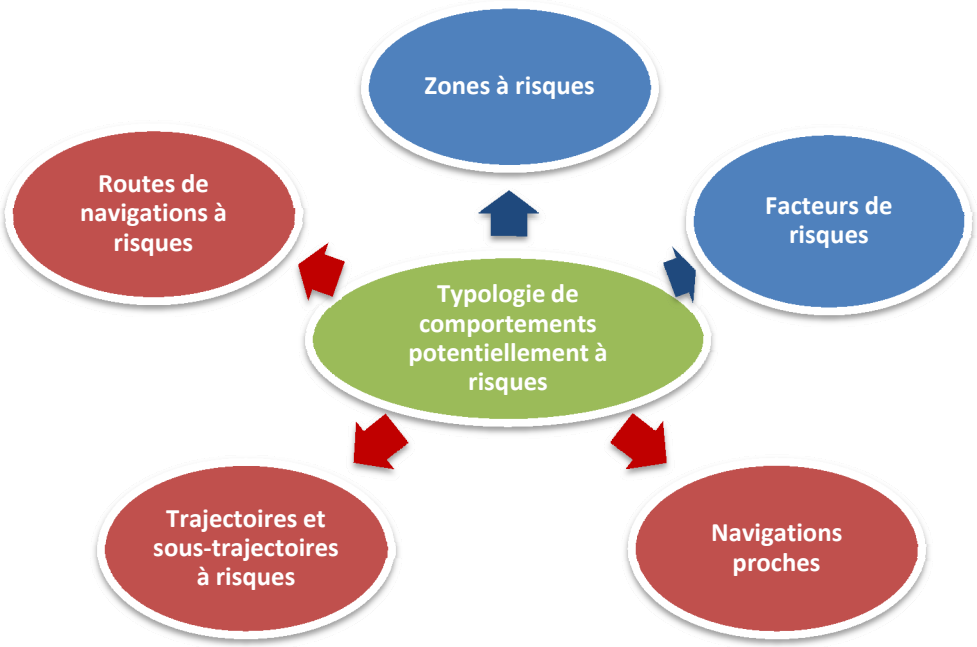


Figure 3-1 : Typologies non exhaustive de comportements potentiellement à risques choisis pour être des fonctionnalités de l'atelier ShipMine.

Cette typologie est composée de cinq comportements :

- **Facteurs à risques :** ce sont des relations entre des facteurs propices à une situation à risque comme par exemple, une relation entre un type de navire et un accident quand il se trouve dans des conditions de navigation particulières,

- **Zones à risques** : zones à forte concentration spatiale d'événements à risques (zones accidentogènes, de piraterie, etc.),

- **Trajectoires et sous trajectoires à risques** : trajectoires décrivant potentiellement un comportement à risque (trajectoire de dérive, trajectoire avec plusieurs changements de directions, etc.),

- **Routes de navigation à risques** : trajectoires type décrivant des navigations à risques comme par exemple les trajectoires de navires transportant des matières dangereuses et sensibles,

- **Navigations proches** : trajectoires proches décrivant les navires qui se rapprochent pendant une durée de temps. Cela peut représenter un risque d'abordage, de transbordement ou de pêche parallèle.

3.2.3. Ciblage et adaptation des algorithmes de fouille de données

Dans les sous-sections suivantes, nous exposons la méthode qui a été adoptée pour identifier, évaluer et tester les algorithmes de fouille de situations et de mouvements pour l'extraction de connaissances décrivant des risques. Nous avons présenté dans la section 3.2.2, quelques comportements à risques. L'extraction de ces comportements est assurée par les fonctionnalités de ShipMine. Il faut trouver pour ces fonctionnalités des méthodes et des algorithmes de fouille de données pouvant les assurer. La Figure 3-2 montre bien cette correspondance entre les fonctionnalités, les méthodes de fouille de données et les algorithmes qui ont été choisis. Nous avons pris la précaution de choisir des méthodes issues de plusieurs domaines, à savoir la fouille de données classique, spatiales et spatio-temporelles pour montrer leur intérêt.

A ce stade de développement, notre atelier ShipMine n'intègre pas toutes les méthodes présentées sur la Figure 3-2 mais juste celles qui ont recours à un affichage cartographique. C'est le cas des méthodes d'analyse de mouvements et l'une des méthodes d'analyse de situations (détection de zones denses). Les méthodes de détection des associations vont être utilisées dans un programme tiers, en dehors de ShipMine. A terme, toutes les méthodes seront intégrées.

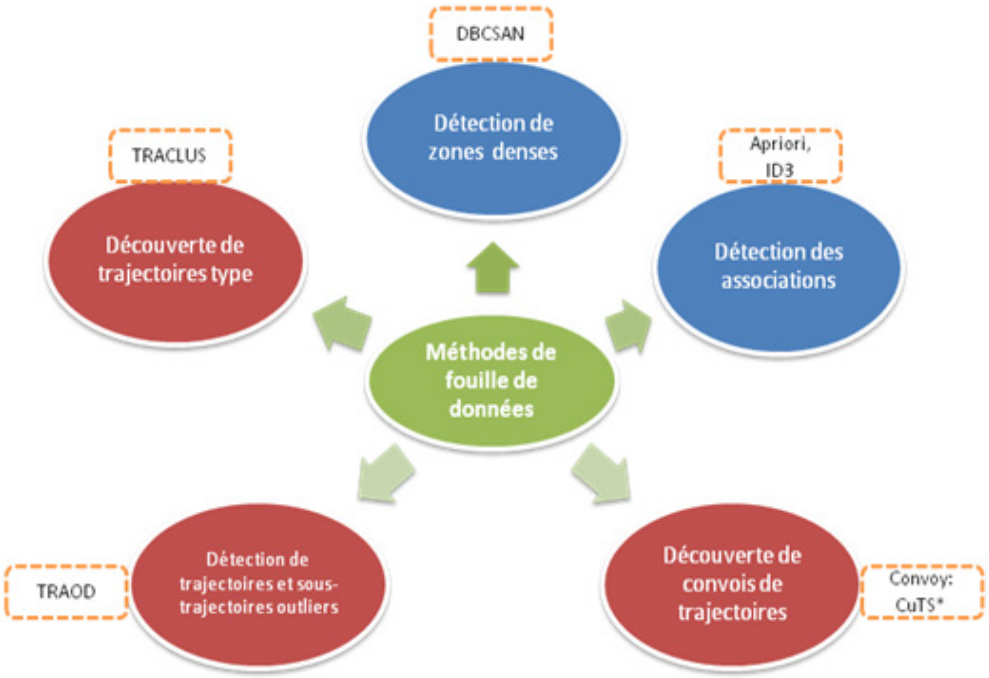


Figure 3-2 : Les méthodes de fouille de données et les algorithmes choisis pour extraire des situations (bleu) et des mouvements (rouge) pouvant décrire des comportements à risques.

3.2.3.1. Analyse de situations

Une situation à risque est définie par une relation entre facteurs propices à un type de risque ou par une concentration spatiale d'événements à risques. Dans l'analyse des relations, nous avons comme ambition de sélectionner des méthodes permettant de découvrir des relations entre des facteurs de contexte (avarie, feu cassé, etc.), d'environnement (conditions météorologiques, océanographiques, etc.) et l'apparition d'événements non souhaités. Dans l'analyse de concentration, nous voulons sélectionner des méthodes permettant de découvrir des zones à risques à partir de concentrations spatiales d'événements à risques.

3.2.3.1.1. Détection de facteurs de risques

Pour découvrir les facteurs de risques et leurs relations automatiquement, il est possible de mettre toutes les variables au même niveau et de chercher les relations (implication, corrélation, etc.) entre elles ou de définir une variable cible à expliquer par d'autres variables dites explicatives. Les deux approches ont été testées et présentées ci-dessous. La première utilise les règles d'association et la deuxième les arbres de décision.

Les règles d'association ont été choisies dans l'objectif de trouver des régularités dans les bases de données d'événements à risques sous forme d'éléments associés fréquemment ensemble. Nous nous sommes inspirés pour cela de l'application du panier de la ménagère (Cf. section 2.2.1.1 du chapitre 2).

Nous avons défini trois grandes catégories de résultats des règles d'association qui sont de la forme « *si antécédent alors conséquent* », pour faciliter leur exploitation :

- **Règles de prédiction** : Nous appelons « règle de prédiction » toute règle ayant son antécédent connu *a priori*, son conséquent non connu et la confiance de la règle est supérieure à 50%. Une règle de prédiction peut être du genre "*Si nous avons un accident dans un contexte Ci alors à c% des cas, il est causé par le type d'accident Ti* ",

- **Règles de ciblage** : Ce sont les règles de connaissances générales qui identifient les relations entre les différentes dimensions (type de navire, type d'accident, zone maritime, etc.). L'antécédent et le conséquent de la règle sont connus mais pas la relation d'implication entre les deux parties. Les règles sont par exemple du genre "*Les accidents de navires de type Ti, concernent à c% les navires de type Ni*" et "*Les accidents de navires de type Ti sont localisés dans c% des cas dans la zone Zi*",

- **Règles banales** : Ce sont les règles qui n'apportent pas d'informations nouvelles.

Pour l'implémentation, l'algorithme *Apriori* (Srikant & Agrawal 1996) développé par *Christian Borgelt* a été choisi. Cette implémentation est récupérée à partir du package *Rattle 2.6.4* de R[70]. Les mesures *support*, *confiance et Lift* (Cf. section 2.2.1.1) vont être utilisées pour sélectionner les règles intéressantes. Le lift est calculé en plus du support et de la confiance pour vérifier que les résultats obtenus ne sont pas le fruit du hasard. Si la mesure du lift est supérieure à 1, la règle est gardée pour une analyse de pertinence.

Dans la deuxième approche, nous allons définir une variable cible à expliquer par d'autres variables dites explicatives. Pour cela, le classement par arbre de décision a été

[70] http://www.inside-r.org/packages/cran/rattle

choisi car il fournit des règles explicites de classement, il est robuste et sa représentation en graphe permet une meilleure facilité de compréhension. L'algorithme ID3 implémenté dans Weka[71] 3.6.8 a été choisi. Cet algorithme est détaillé dans la section 2.2.1.2.

3.2.3.1.2. Détection de zones à risques

Le groupement de localisations d'événements non souhaités comme les accidents de navires en utilisant une mesure de proximité géographique va peut être nous permettre d'extraire des zones à risques. Pour découvrir ces zones à risques, plusieurs méthodes de fouille spatiale et plus particulièrement de clustering peuvent être utilisées comme les méthodes par partitionnement, les méthodes hiérarchiques et les méthodes par densité (Cf. section 2.2.1.3).

Nous avons choisi d'utiliser les méthodes de clustering par densité pour deux raisons. La première concerne la détection automatique du nombre de clusters ce qui permet de ne pas biaiser les résultats en fixant ce nombre *a priori*. La deuxième raison est relative à la capacité de ce type de méthode à récupérer des clusters ayant des formes arbitraires. Parmi les algorithmes de clustering de densité, notre choix s'est porté sur DBSCAN par ce qu'il présente une faible complexité *O(n log n)* et les résultats générés sont visuels et textuels ce qui permet une validation plus facile de ces résultats. Pour plus de détails sur cet algorithme voir la section 2.2.1.3.

Pour l'implémentation de DBSCAN, un programme *java* est proposé sur internet[72] en téléchargement et exploitation libre. Ce programme est proposé avec une Graphical User Interface (GUI) que nous avons dû enlever pour pouvoir appeler le programme d'une manière transparente à partir de ShipMine. Nous avons réalisé d'autres modifications pour retourner les résultats à ShipMine à la fin de l'exécution et pour calculer des polygones englobant les zones de densité découvertes. L'appel du programme à partir de ShipMine comprend les valeurs des paramètres d'Eps (rayon de voisinage), MinPts (seuil minimum de points), les chemins des fichiers de données à explorer et ceux des résultats de l'exécution.

[71] http://www.cs.waikato.ac.nz/ml/index.html
[72] https://github.com/KanwarBhajneek/DBSCAN/tree/master/src/dbscan

Les résultats retournés sont de deux types : les clusters avec les positions d'événements associées et les polygones englobant chaque cluster. Les polygones sont calculés en utilisant le parcours de Graham (Graham & Yao 1983) pour représenter les zones à risques. Le parcours de Graham décrit les plus petits polygones contenant l'ensemble des points de chaque cluster. Le code source de cet algorithme est proposé en téléchargement libre[73].

3.2.3.2. Analyse de mouvements à risques

Les mouvements potentiellement à risques peuvent êtres des formes de trajectoires particulières, la proximité de deux ou plusieurs trajectoires et des formes de trajectoires différentes de l'habituel (trajectoire avec des cercles, des zigzags, etc.). Ces mouvements peuvent décrire un risque comme une dérive, une pollution, un objet tombé à la mer ou un accident. Pour extraire ces motifs, nous nous sommes inspirés des travaux de l'équipe de J. Han (Lee et al., 2007) (Lee et al., 2008a) (Li et al., 2010c) qui proposent l'identification de trajectoires aberrantes (*Outlier Trajectory*), le clustering de trajectoires et les travaux de Jeung sur les convois (Jeung et al., 2008a; Jeung et al., 2008b).

3.2.3.2.1. Détection de trajectoires anormales de navires

Pour extraire les motifs de trajectoires anormales peuvant correspondre à un risque, l'algorithme d'extraction de trajectoires aberrantes TRAOD (Cf. section 2.2.3.1.1 du chapitre 2) a été choisi. Nous rappelons qu'une trajectoire aberrante est une trajectoire qui ne ressemble pas au comportement général des autres trajectoires voisines.

Le programme Visual C++ de cet algorithme a été téléchargé à partir de la page personnelle de Jae-Gil Lee[74] qui est l'un des auteurs de TRAOD. Ensuite, il a été testé sur leurs données d'expérimentation pour vérifier qu'il donne les mêmes résultats.

Le programme original possède une GUI pour sélectionner les données, saisir les paramètres, exécuter le programme et afficher les résultats sur une interface graphique. Les auteurs ont utilisé un référentiel spatial relatif à l'interface d'affichage.

[73] Code source sur http://www.script-tout-fait.com/fr/script-Concevoir-une-enveloppe-convexe-avec-l-algorithme-de-Graham-31.html
[74] http://dm.kaist.ac.kr/jaegil/sources/

Nous avons enlevé l'interface du programme pour permettre son exécution d'une manière transparente par rapport à l'utilisateur, qui aura l'impression que tout se fait au niveau de ShipMine. D'autres changements ont été opérés pour adapter le programme à notre application comme la modification des paramètres en entrée, omettre la phase de simplification des données MDL (Cf. section 2.2.3.1.1) et la repondération des calculs de distance. Ces changements sont détaillés ci-après.

Le paramètre F représentant la proportion de longueur des partitions aberrantes pour que toute la trajectoire soit considérée comme aberrante a été écarté. Le fait que les trajectoires soient très longues, la proportion F même petite soit-elle, élimine beaucoup trop de trajectoires aberrantes. L'idée est de garder toutes les partitions aberrantes quelle que soit la proportion de ses partitions sur l'ensemble de la trajectoire.

Après avoir fait plusieurs tests sur nos jeux de données, nous nous sommes rendus compte que l'algorithme d'origine ne détectait que les longues partitions dues souvent à la perte de signal AIS et non à des comportements de navires aberrants. La Figure 3-3 montre l'un des résultats de ces tests qui a donné 30 pertes de signal AIS à partir d'un historique de 65 trajectoires de tankers localisés en Méditerranée. Le résultat obtenu n'est pas celui espéré car l'objectif est d'identifier les partitions aberrantes des trajectoires dues à des mouvements de navires et non à des pertes de signal AIS. Tout porte à croire que la simplification MDL utilisée, enlève les petites partions aberrantes qui nous intéressent, c'est pour cela qu'elles ne sont pas détectées. En effet, après la simplification, les trajectoires sont devenues trop lisses, peut être à cause de l'hétérogénéité des formes et des longueurs de partitions de trajectoires.

Nous avons choisi de suspendre la simplification MDL pour remédier au problème. Ce qui donne une partition de trajectoire entre deux positions successives. Cela n'aura pas d'impact sur les performances car nous travaillons sur des données non volumineuses.

Figure 3-3 : Partitions de trajectoires aberrantes détectées après une phase de simplification MDL (p=98%, D=10Km).

Les distances peuvent être pondérées selon l'importance que l'on veut donner à chacune des distances que sont les distances perpendiculaire, parallèle et angulaire (Cf. section 2.2.3.1.1, page 84). Après plusieurs tests, les pondérations ont été fixées à $w_{\perp}=1$, $w_{\parallel}=0$, $w_{\theta}=5$ car cela donne plus d'importance à la distance angulaire et permet de détecter les trajectoires qui font par exemple des zigzags, des cercles et des dérives. La distance parallèle a été mise à zéro car elle n'a pas d'intérêt dans notre application dû au fait que les positions ne sont pas récupérées d'une manière synchrone.

3.2.3.2.2. Découverte de routes de navigation

Dans la majorité des cas, les navires de transport de personnes et de marchandises suivent les mêmes routes. Ils font souvent des allers et retours entre différents ports. La découverte de ces routes de navigation peut aider par exemple à identifier les routes de transport de matières dangereuses, polluantes ou identifier les trajectoires de navires qui ne suivent pas ces routes de navigation (comportement anormal).

Plusieurs méthodes de groupement de trajectoires peuvent être utilisées pour la découverte de ces routes. Notre choix s'est porté sur les méthodes de clustering de trajectoires basées sur les densités. L'algorithme TRACLUS vu dans la section 2.2.3.1.2 répond bien à nos attentes. Il est basé sur l'algorithme DBSCAN. Il utilise donc les mêmes concepts de densité en considérant les segments de ligne à la place des points. Une trajectoire représentative est créée pour chaque cluster, cette trajectoire décrit le mouvement du cluster.

L'algorithme commence par partitionner les trajectoires en petites partions avant de commencer la phase de groupement. Dans le partitionnement, il utilise des points caractéristiques pour simplifier les trajectoires avant leur exploration. Le choix de l'algorithme est motivé par le fait qu'il soit précis et efficace.

Les auteurs de TRACLUS proposent le téléchargement du code source C++ de l'algorithme sur internet[75]. Des modifications ont été apportées au programme pour pouvoir l'appeler à partir de ShipMine d'une manière transparente aux utilisateurs.

Après plusieurs tests sur la pondération de distances utilisée dans TRACLUS, elle a été fixée d'une manière empirique à $w_{\perp}=1$, $w_{\parallel}=0$, $w_{\theta}=1.2$. Cette pondération donne un peu plus d'importance à la distance angulaire et donne de meilleurs résultats. La distance parallèle a été mise à zéro car elle n'a pas d'intérêt du fait que les positions ne sont pas récupérées d'une manière synchrone. Cette pondération permet à l'algorithme de regrouper les partitions de trajectoires parallèles car elles sont considérées comme plus proches (voisines) que celles ayant des différences angulaires.

Le calcul de la trajectoire représentative à partir des clusters de segments de lignes est basé sur le concept de connexion de densité (Cf. section 2.2.3.1.2). Pour éviter de connecter les clusters espacés entre eux, une valeur de la distance égale à 500 mètres a été choisie après plusieurs tests. La validation de cette distance a été effectuée d'une manière visuelle en comparant les différents résultats de tests.

3.2.3.2.3. Découverte de navigations proches et parallèles

Trouver les groupes de navires qui se déplacent ensemble dans un voisinage proche peut être intéressant pour la découverte de risques de pêche parallèle, d'abordage ou de trafic illicite. Des motifs spatio-temporels de trajectoires proches et parallèles peuvent permettre la découverte de ces groupes de navires et leur durée de déplacement ensembles. Ces motifs sont des clusters spatiaux reliés successivement pendant une durée de temps par un seuil d'objets en communs. Plusieurs méthodes de clustering d'objets mobiles existent dans la littérature comme Swarm, Flock et Convoy (Cf. section 2.2.3.1.4 du chapitre 2).

[75] http://dm.kaist.ac.kr/jaegil/sources/

Notre choix s'est porté sur l'algorithme Convoy par ce qu'il permet de détecter les objets qui se déplacent ensemble pendant une durée de temps minimale. Cela est en adéquation avec notre problématique du fait qu'il est impossible de savoir par avance les durées de temps passées par des navires à naviguer à proximité. De plus, il se base sur la notion de densité ce qui nous permet de détecter des convois ayant des formes arbitraires.

Le programme de l'algorithme Convoy a été mis à notre disposition par J. Hoyoung qui est l'un de ses auteurs. Des modifications ont été apportées à cette implémentation pour l'intégrer à ShipMine. Les valeurs des paramètres m, k, e, δ (Cf. section 2.2.3.1.4) ainsi que le chemin des données à explorer et ceux des résultats sont fournis en argument de l'appel du programme. Après exécution, L'algorithme Convoy génère trois fichiers. Le premier, contient des métriques sur l'exécution à savoir, les valeurs de paramètres, le temps d'exécution, les convois candidats et les convois actuels. Le deuxième fichier comporte les convois actuels et le troisième, les trajectoires simplifiées par l'algorithme DP* (Cf. section 2.2.3.1.4). Le deuxième fichier est exploité par ShipMine pour afficher les convois résultats.

3.2.4. Structure de l'espace de données à explorer

Dans une approche de fouille de données, une phase d'acquisition des données à explorer est nécessaire. Les données à explorer sont de deux catégories : les données spatiales statiques et spatiales dynamiques.

3.2.4.1. Les données spatiales statiques

Ce sont des données décrivant des événements ponctuels à risques qui se sont produits à une date et à une localisation données. Ces données sont structurées comme suit :

<Événement (var_1,…,var_n), Timestamp, Localisation (Latitude, Longitude)>

- *Événement (var_1,…,var_n)* : Décrit un événement par un ensemble de variable comme le type d'événement, les objets concernés, le contexte et l'environnement,
- *Timestamp* : La date et l'heure à laquelle l'événement s'est réalisé,
- *Localisation (Latitude, Longitude)* : la localisation absolue de l'événement.

Ces données peuvent être par exemple des données décrivant des accidents maritimes ou des actes de piraterie.

L'exploration de cette catégorie de données, peut permettre la découverte de connaissances sur des situations à risques qui peuvent être soit des relations entre les valeurs, des variables décrivant un événement, ou des zones à forte densité des événements étudiés.

3.2.4.2. Les données spatiales dynamiques

Ce sont des données historiques recensant les déplacements d'objets. La structure de ces données est définie comme suit :

<Objet (var_1,…,var_n), Localisation <loc1,…,locm>, Timestamp <t1,…tm >>,

- *Objet (var_1,…,var_n)* : Décrit un objet mobile par un ensemble de variable comme le type d'objet, sa vitesse et l'environnement d'évolution,
- *Timestamp* : La date et l'heure à laquelle la donnée a été acquise,
- *Localisation (Latitude, Longitude)* : La localisation absolue de l'objet à l'instant *Timestamp*.

Ces données peuvent être des traces AIS ou radar. L'exploration de cette catégorie de données par des méthodes de fouille de données d'objets mobiles peut permettre l'extraction de motifs de mouvements décrivant des comportements à risques.

Les données spatiales statiques et dynamiques qui ont été acquises pour le test et la validation de notre méthode sont présentées dans le chapitre suivant (Cf. section 4.2.1).

3.2.5. Architecture

Dans notre architecture représentée sur la Figure 3-4, trois interfaces sont distinguées : une interface de visualisation ; une interface de fouille de données et ; une interface des données à explorer. Cette architecture a été pensée d'une manière extensible pour pouvoir accueillir d'autres fonctionnalités.

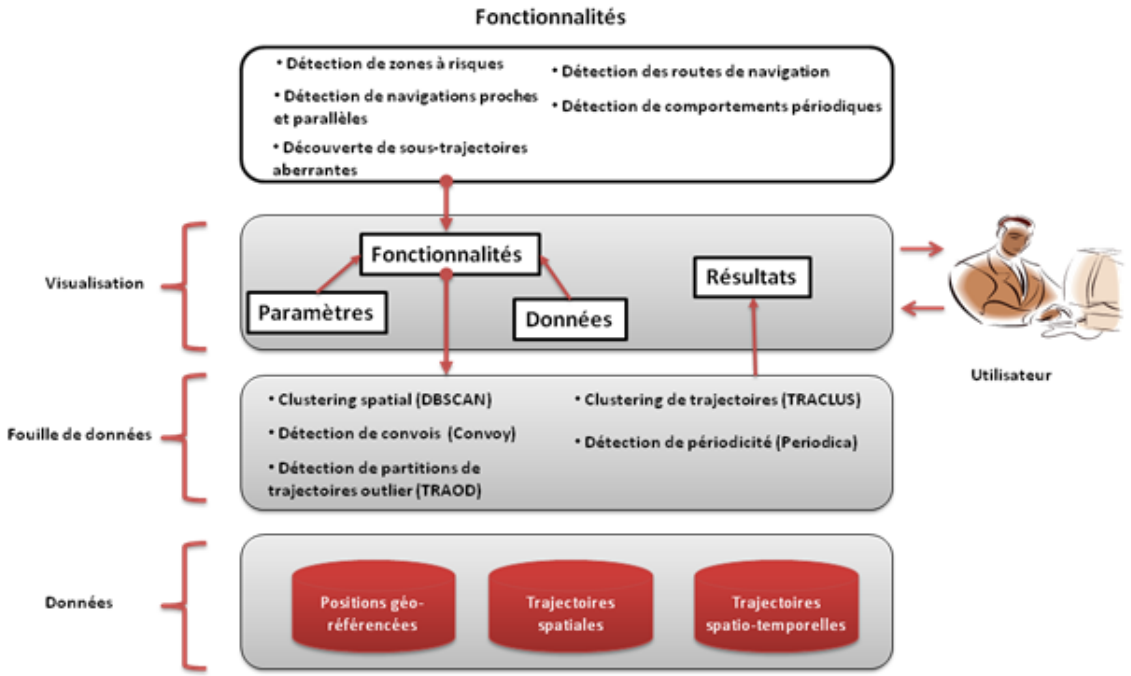


Figure 3-4 : Architecture système de ShipMine.

L'architecture de ShipMine à été inspirée de l'architecture proposée dans MoveMine (Li et al., 2011).

3.2.5.1. Interface de visualisation

L'utilisateur de l'atelier peut interagir avec l'interface de visualisation pour choisir une fonctionnalité, sélectionner des données, visualiser ces données sur une cartographie, entrer les valeurs des paramètres, exécuter une fonctionnalité et interagir avec les résultats affichés sur la cartographie (sélectionner, zoomer, afficher des informations complémentaires, etc.).

Pour comprendre le fonctionnement de ShipMine et ses interactions avec l'utilisateur, un diagramme de cas d'utilisation UML[76] de l'interface *Visualisation* est représenté sur la Figure 3-5. Ce diagramme montre une vue générale et simplifiée des cas d'utilisations de ShipMine et leur relations d'inclusion et d'extension

.

[76] Unified Modeling Language est un langage de modélisation graphique utilisé dans le développement logiciel.

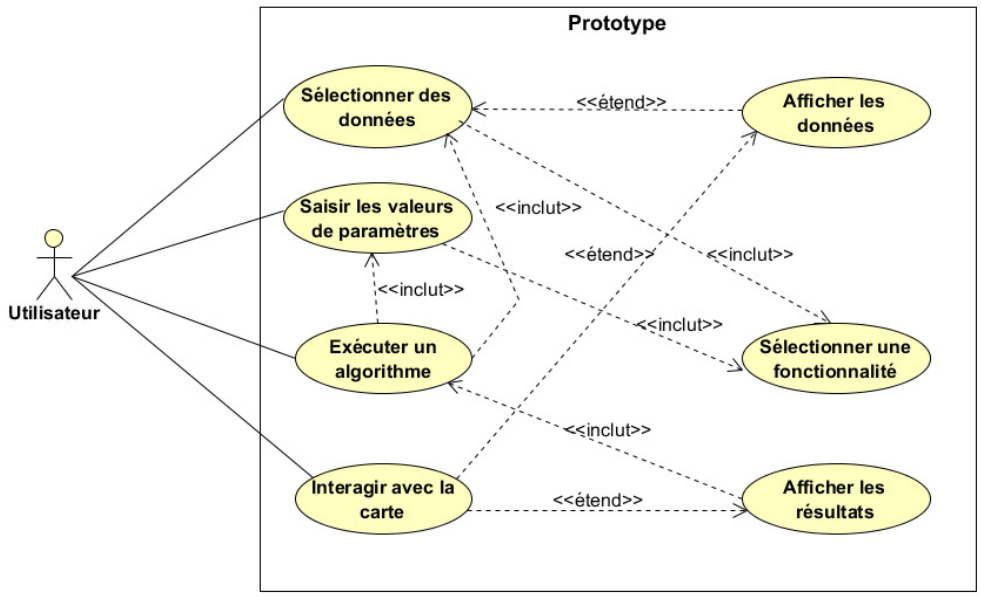


Figure 3-5 : Diagramme de cas d'utilisation de l'interface Visualisation de ShipMine.

Il est à noter sur ce diagramme (Figure 3-5) que le cas d'utilisation *Sélectionner une fonctionnalité* inclut forcément celui de la sélection des données et la saisie des valeurs des paramètres. Ces deux cas à savoir, *Sélectionner des données* et *Saisir les valeurs de paramètres* sont inclus à leur tour dans le cas *Exécuter un algorithme*. Concernant le cas d'utilisation *Interagir avec la carte*, l'utilisateur peut optionnellement afficher les données et/ou les résultats pour interagir avec eux.

L'interface *Visualisation* est l'entrée du système. C'est à partir de cette interface que l'utilisateur peut communiquer avec les composantes du système, se trouvant dans différentes interfaces.

3.2.5.2. Interface de fouille de données

L'interface *Fouille de données* est le cœur du système. Elle contient un ensemble d'algorithmes implémentés, adaptés et intégrés au système. Quand une fonctionnalité est choisie par l'utilisateur, cette interface est sollicitée pour exécuter l'algorithme associé.

L'exécution d'un algorithme dans ShipMine passe par une séquence chronologique d'actions. Cet enchaînement d'actions est représenté sur la

Figure 3-6 par un diagramme de séquence UML.

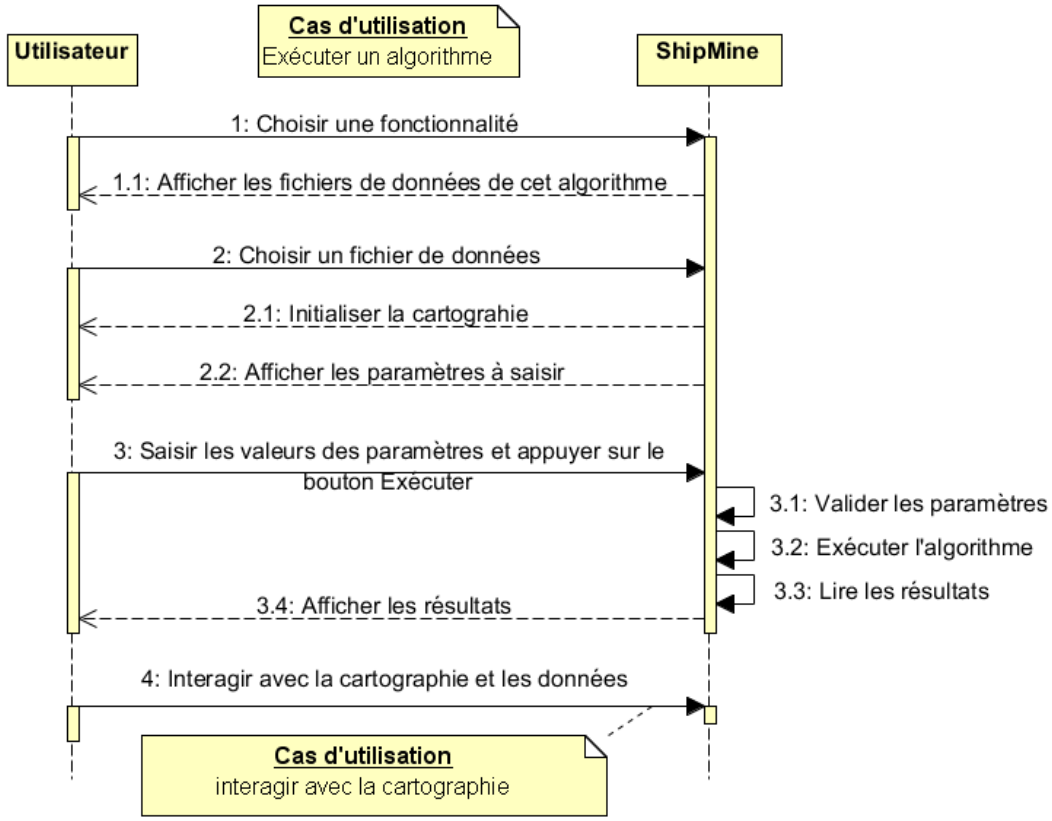


Figure 3-6 : Diagramme de séquence des interactions au cours de l'exécution d'un algorithme.

Quand l'utilisateur choisit une fonctionnalité, ShipMine lui associe un algorithme et affiche une liste de jeux de données spécifiques à cet algorithme. Chaque algorithme utilise un format particulier des données en entrée.

L'utilisateur choisit un jeu de données à explorer et le système initialise la cartographie, affiche le contenu de ce jeu de données et les champs des paramètres à saisir. L'utilisateur est libre d'interagir avec ce jeu de données initiales (zoomer, afficher des informations complémentaires, etc.) ou exécuter la fonctionnalité. Si le bouton *Exécuter* est pressé, le système valide les valeurs des paramètres saisies, appelle le

programme associé et affiche les résultats de l'exécution pour que l'utilisateur puisse les étudier en interagissant avec la cartographie.

La représentation d'un cas d'utilisation d'une interaction avec la cartographie est présentée par un diagramme de séquences UML (Figure 3-7). La première séquence du diagramme représente la réponse de l'API[77] Google Maps aux actions de l'utilisateur comme le clic, le double clic et le changement de type de cartographie (plan, satellite, relief, etc.). Ces actions sont implantées par défaut dans l'API Google Maps qui va être présentée plus en détail dans la section 3.3.

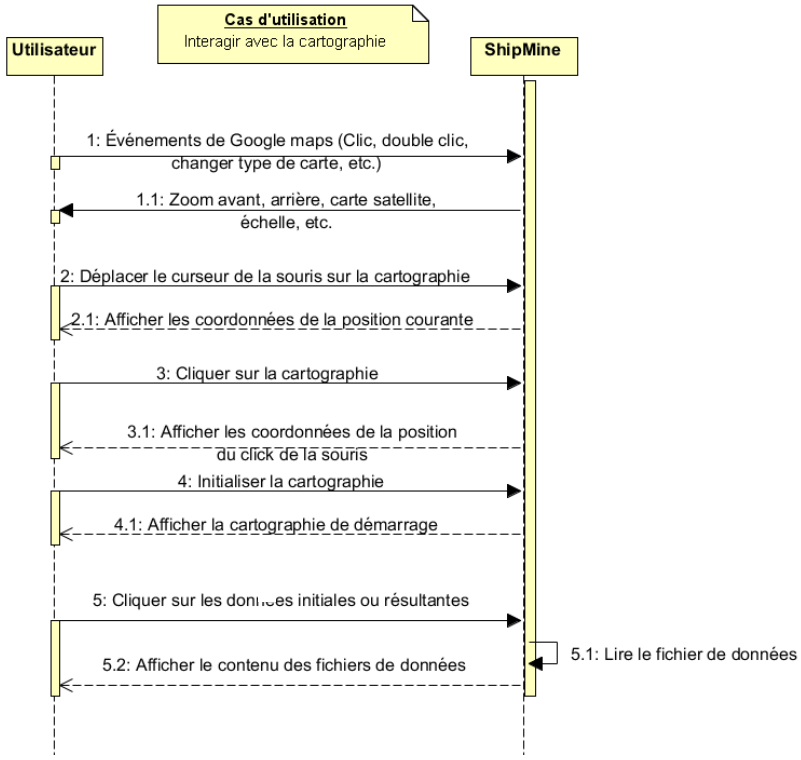


Figure 3-7 : Diagramme de séquence des interactions de l'utilisateur avec la cartographie.

En plus des actions implantées par défaut dans l'API Google Maps, d'autres sont ajoutées à ShipMine concernant l'affichage de coordonnées géographiques du déplacement de la souris, l'affichage de coordonnées de positions cliquées, l'affichage de nos données, de nos motifs sur cette cartographie ainsi que d'autres informations attributaires.

[77] Application Programming Interface est une interface de programmation contenant un ensemble de classes, méthodes et fonctions qu'il est possible d'utiliser dans de nouveaux programmes.

3.2.3.3. Interface de données à explorer

Concernant l'interface de données de l'architecture ShipMine, elle comporte des données sur des déplacements de navires et des événements maritimes passés. Ces données sont récupérées, traitées et transformées selon l'exigence des algorithmes de fouille de données qui vont les utiliser. Chaque algorithme peut avoir un format et un contenu de données spécifiques.

Dans cette interface, il est possible de distinguer trois types de données : positions géo-référencées, les trajectoires spatiales et les trajectoires spatio-temporelles. Les trajectoires spatiales sont composées de séquences de positions géo-référencées et les trajectoires spatio-temporelles sont des trajectoires spatiales où le temps de chaque position est renseigné.

Cette typologie identifie la nature des données et par conséquent les méthodes de fouille de données qui peuvent être utilisées.

3.2.6. Choix technologique

Nous avons opté pour une application web car elle permet d'avoir une architecture client-serveur avec un accès client léger. Le développement de l'application a été réalisé lors d'un stage que nous avons encadré (El Moussawi, 2013). L'utilisateur n'a rien à installer sur son ordinateur et il n'a besoin que de son navigateur web pour envoyer des requêtes au serveur et afficher les résultats sous forme de pages web. Toutes les interactions avec le navigateur sont transformées en requêtes. L'atelier est composé de pages web stockées sur un serveur web de type Apache et jouant le rôle d'interface entre le client et les implémentations des algorithmes. Les échanges entre le client et le serveur se font par le protocole HTTP. Nous n'avons pas utilisé un protocole sécurisé car les données qui circulent entre le client et le serveur sont rendues anonymes et peuvent donc transiter en claire sur le réseau.

Pour représenter les pages web, le langage HyperText Markup Language (HTML) est utilisé. L'acronyme HTML signifie que le langage permet de traiter des hypertextes en se basant sur un langage de balisage. Les balises permettent de mettre en forme les textes affichés sur la page web et d'afficher des éléments interactifs comme les boutons, les images et les liens hypertextes. Pour gérer les interactions côté client et avoir des pages

dynamiques, le langage de programmation JavaScript est utilisé. Ce langage apporte des améliorations à l'HTML en permettant d'exécuter des commandes au niveau client, c'est-à-dire par le navigateur web du client. L'affichage des résultats d'exécution des algorithmes de l'atelier se fait en utilisant JavaScript. Dans le développement de ShipMine, nous avons choisi d'utiliser le PHP pour l'appel des exécutables binaires des algorithmes qui composent l'atelier et lire les résultats des exécutions à partir de fichiers. Le PHP[78] est un langage de programmation très utilisé pour la création de pages web dynamiques. Ce qui a motivé ce choix est le fait qu'il permette au serveur de répondre au client sans dévoiler le contenu du code source exécuté. De plus, la version 5 de de PHP utilise la programmation orientée objet ce qui permet une réutilisation plus souple des classes d'objets.

Une autre technologie utilisée comme partie intégrante de l'IHM (Interface Homme Machine) de ShipMine est l'API Google Maps. Cette API permet de gérer l'affichage et la manipulation des objets cartographiques. Elle contient une bibliothèque de fonctions et de classes permettant de gérer des objets géographiques et des rasters dynamiquement à partir d'une interface web. Les services de base intégrés dans cette API sont par exemple le contrôle du zoom, le déplacement dans la carte, le changement d'échelle, le changement du type de carte, l'affichage des info-bulles, l'affichage des objets géométriques (positions, polylignes, polygones), le contrôle des événements ou des actions des utilisateurs sur la carte (superposition de plusieurs couches, etc). Les positions des navires vont être représentées dans Google Maps par des marqueurs, les trajectoires par des polylignes et les zones par des polygones. Les informations complémentaires sur les objets cartographiques vont être affichées dans des info-bulles.

Les programmes qui ont été choisis pour intégrer notre atelier sont développés la plupart du temps en Visual C++, C++ et Java. Le choix de ces technologies de développement s'est imposé par souci de réutilisation des codes sources existants.

3.3. Présentation de ShipMine

Après avoir sélectionné les algorithmes et préparé les données à utiliser, leur intégration dans l'atelier ShipMine a été opérée de manière à pouvoir ajouter d'autres

[78] http://php.net/docs.php

algorithmes et d'autres données à notre atelier. Les fichiers de données qui peuvent être analysés sont ceux déjà intégrés au préalable au système. Par souci d'intégrité des données, l'utilisateur de ShipMine ne peut pas analyser d'autres données sauf si l'administrateur les intègre dans l'un des répertoires de données créés à cet effet et paramètre le système pour qu'il prenne en compte ces données.

Comme les données brutes, les motifs résultants peuvent être aussi représentés par une typologie contenant trois formes géométriques (Figure 3-8) : un point, un polygone et une polyligne.

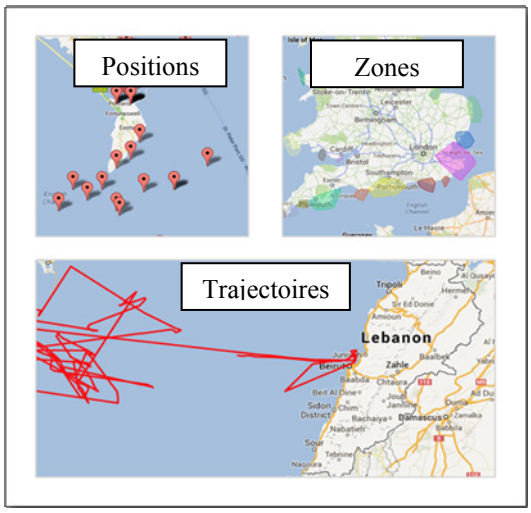


Figure 3-8 : Typologie de formes géométriques utilisée pour la représentation des données et des motifs dans ShipMine.

L'interface de ShipMine est composée de quatre cadres comme on le voit sur la Figure 3-9. Le premier est l'entête, il contient la bannière ; le deuxième est composé de deux listes déroulantes permettant de choisir quelles fonctionnalités et données utiliser ; le troisième cadre quant à lui concerne la saisie des valeurs de paramètres et le lancement de l'extraction et ; le quatrième et dernier cadre est l'interface cartographique Google Maps permettant d'afficher et d'interagir avec les données.

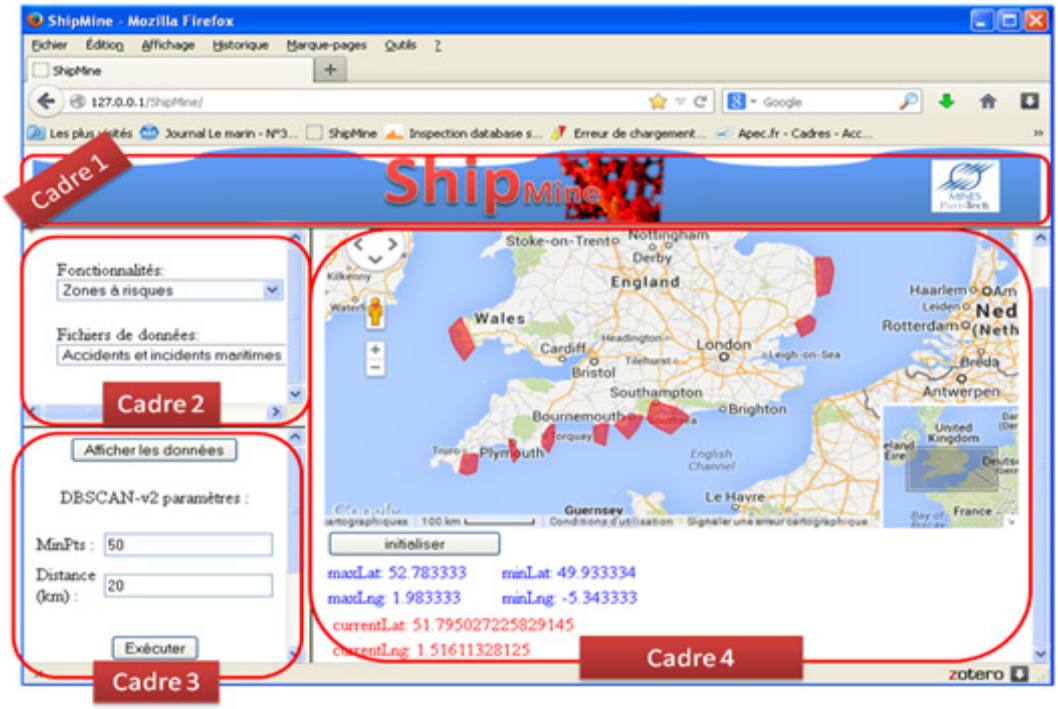


Figure 3-9 : Interface principale de ShipMine avec ses différents cadres.

Les différents cadres de l'interface sont gérés par des fichiers HTML et PHP permettant de faire des tâches spécifiques. Le fichier index est le premier fichier qui s'exécute pour initialiser l'interface. Ce fichier appelle par la suite, les autres fichiers qui sont le Header.html, Algorithme-data.php, Paramètres.php et Cartographie.php. Nous allons détailler par la suite chaque cadre de notre interface.

3.3.1. Bannière

Dans ce cadre, la bannière graphique est chargée à partir d'une image présentant le nom de l'atelier ainsi que le logo de MINES ParisTech.

3.3.2. Choix de la fonctionnalité et des données

Ce cadre est constitué de deux listes déroulantes. La première liste l'ensemble des fonctionnalités qu'il est possible d'utiliser et la deuxième les données qui sont liées à la fonctionnalité choisie (Figure 3-10). Ces listes sont créées dynamiquement au cours de l'initialisation de l'application en se basant sur un fichier de paramètres. Cela permet à l'atelier de rester évolutif en gardant la possibilité de changer la liste des fonctionnalités

en enlevant ou ajoutant d'autres fonctionnalités sans toucher au code source. La liste des données quant à elle est récupérée à partir du répertoire de données spécifié dans le fichier paramètres appelé « algorithmes implémentes.txt ». Pour ajouter une autre fonctionnalité, il suffit de rajouter une ligne dans ce fichier contenant, le nom de la fonctionnalité, les paramètres de l'algorithme associé à la fonctionnalité et le chemin du répertoire contenant les données.

Figure 3-10 : Cadre du choix d'une fonctionnalité et des données.

3.3.3. Fonctionnalité d'exécution et de paramétrage

Dans ce cadre d'interface, il est possible d'afficher les contenus des fichiers de données, de saisir les paramètres et d'exécuter l'exploration des données (voir Figure 3-11). Le programme associé à la fonctionnalité sélectionnée est exécutée avec les valeurs des paramètres transmises par le système. Des contrôles de vérification et de validation des paramètres saisis sont intégrés au système. Un message d'erreur est retourné si les paramètres saisis sont erronés. Prenons l'exemple du paramètre minimum de positions pour former un cluster (MinPts) de l'algorithme DBSCAN (Cf. section 2.2.1.3). La valeur de ce paramètre doit impérativement être entière et supérieure à deux.

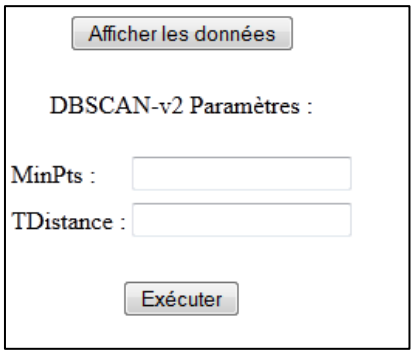


Figure 3-11 : Cadre paramètres et exécution de l'exploration.

Le cadre *Paramètres* dépend de la fonctionnalité choisie. En effet, avant de choisir la fonctionnalité, le cadre est vide. Quand l'utilisateur sélectionne une fonctionnalité, le système lit le fichier « algorithmes implémentes.txt » (Cf. section 3.3.2) pour récupérer dynamiquement les noms des paramètres associés à la fonctionnalité.

Les résultats de l'exploration sont issus d'un fichier d'échange entre ShipMine et le programme exécuté. Ces résultats vont être affichés dans le cadre de l'interface cartographique comme nous allons le voir dans la section suivante.

3.3.4. Interface cartographique

Ce cadre intègre une interface cartographique Google Maps utilisée pour l'affichage des données, des résultats et l'interaction entre eux. Cette interface cartographique est liée aux autres cadres et les changements effectués par l'utilisateur, les données en entrée ou les paramètres sont visualisés dans ce dernier cadre. Il est possible d'interagir avec les données et les motifs affichés et de donner à l'utilisateur des détails sur les éléments cliqués. Prenons deux exemples d'affichage d'informations, le premier est un clic sur une trajectoire pour afficher un info-bulle indiquant le numéro MMSI du navire et le nombre de positions de la trajectoire (Figure 3-12-a) et le second exemple est un info-bulle sur la localisation de la position et le type d'accident (Figure 3-12-b).

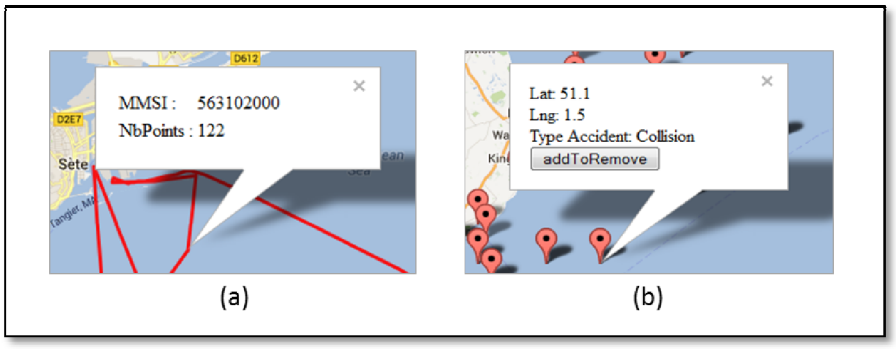


Figure 3-12 : Affichage d'informations attributaires sur des éléments de données (a) trajectoire (b) position.

Deux composants ont été ajoutés sur le bas du cadre cartographique pour permettre l'initialisation de la carte et la récupération d'informations de localisation comme la position du curseur et la position d'un clic de souris sur la cartographie. Les valeurs minimales et maximales de la latitude et longitude (*minLng* et *maxLng*) sont récupérées à partir du fichier de données et affichées dans le même cadre avec une couleur bleu. Les latitudes et longitudes courantes (*currentLat* et *currentLng*) dues au déplacement de la souris sur la carte sont par contre affichées en rouge. Les coordonnées affichées en vert représentent la position du dernier clic de souris sur la cartographie (*clickedLat* et *clickedLng*). Toutes ces petites fonctionnalités vont permettre d'analyser les données et les résultats affichés sur la cartographie.

L'initialisation est accessible à l'utilisateur par le biais du bouton *« Initialiser »*. Comme on le voit sur la Figure 3-13, un autre bouton a été ajouté pour permettre de

supprimer des positions dans le fichier de données originales. A la fin du nettoyage, l'utilisateur peut sauvegarder les données restantes dans un nouveau fichier. Le fichier est mis dans le même répertoire que le fichier original avec une dénomination clean devant l'ancien nom du fichier.

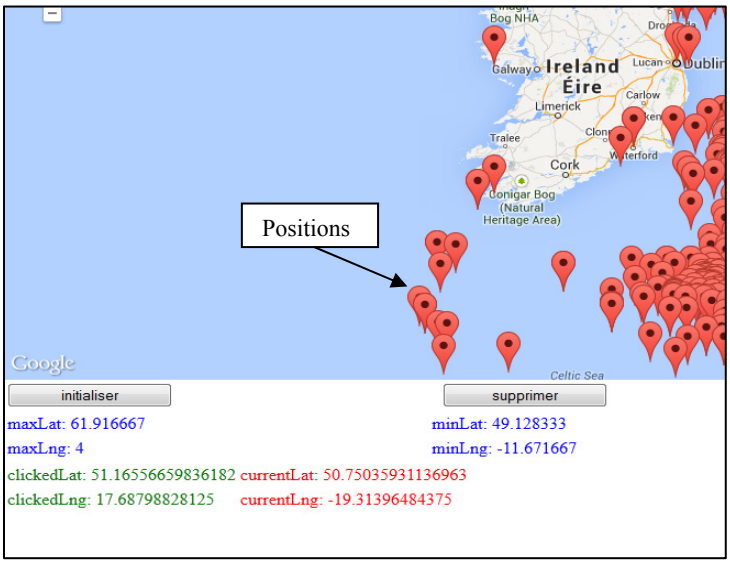


Figure 3-13 : Cadre cartographique de l'interface de ShipMine.

3.4. Conclusion

Dans l'objectif de valider notre méthodologie d'aide à l'analyse des comportements à risques, nous avons identifié et testé des méthodes de fouille de données susceptibles de décrire ces comportements. La fouille de données n'a pas été conçue pour extraire les comportements à risques mais elle permet d'extraire des motifs de mouvements, de situations (anormaux, fréquents, etc.) ou leurs combinaisons qu'il est possible d'interpréter comme étant à risques.

Des méthodes de fouille de données pouvant extraire des situations et mouvements à risques ont été intégrées dans un seul environnement pour servir d'atelier d'extraction de connaissances sur les comportements potentiellement à risques. Les méthodes qui ont été choisies pour extraire ces connaissances sont la détection des associations, la détection des zones à risques, la découverte de convois de trajectoires, la détection de trajectoires

aberrantes et la découverte de trajectoires type. Les étapes suivies pour concevoir et réaliser cet atelier (ShipMine) ont été présentées dans ce chapitre.

Pour tester et valider nos méthodes dans ce contexte d'analyse de comportements à risques, des motifs et règles décrivant les comportements à risques doivent être extraits en utilisant ShipMine. Les résultats de cette exploration vont être présentés et discutés dans le chapitre suivant.

Chapitre 4 : Exemples d’extraction de connaissances sur les comportements de navires potentiellement à risques

4.1. Introduction

Nous allons présenter dans ce chapitre quelques exemples d'extraction de connaissances pour l'analyse de comportements de navires potentiellement à risques dans l'objectif de tester et valider notre méthodologie. L'interprétation des connaissances générées à l'aide de ShipMine sous forme de mouvements et de situations permet de qualifier un comportement comme étant à risque.

Nous avons choisi d'extraire des exemples de comportements potentiellement à risques par type de comportements présentés sur la Figure 3-1 (Cf. section 3.2.2 du chapitre 3) :

- Des associations entre facteurs d'accidents pour le type « Facteurs de risques »,
- Des zones accidentogènes pour « Zones à risques »,
- Des trajectoires de dérives pour « Trajectoires et sous-trajectoires à risques »,
- Des comportements d'abordage et de pêche parallèle pour « Navigations proches »,
- Des routes à risques de naufrage pour « Routes de navigation à risques ».

Pour réaliser ces extractions, nous avons fait l'acquisition de bases de données sur des accidents maritimes et sur les pistes AIS de déplacements de navires qui sont présentées ci-dessous. La préparation de ces données, à savoir leur nettoyage et la restructuration de l'espace de données à explorer par les méthodes de fouille de données, a été aussi exposée. Le nettoyage de ces données permet d'enlever les bruits, les incohérences et traiter les valeurs manquantes.

Ce chapitre est organisé en trois parties : la première partie concerne la préparation de l'espace de données à explorer ; la deuxième s'intéresse à l'extraction de comportements pouvant décrire des comportements à risques et la troisième traite des limites et améliorations de l'approche de validation et des méthodes de fouille de données utilisées. La première partie de ce chapitre est subdivisée en trois sous-parties, à savoir l'acquisition des bases de données, leur nettoyage et la modélisation des espaces de données à explorer. La deuxième partie quant à elle, est subdivisée en deux sous-parties

pour distinguer l'extraction de situation à risques de l'extraction de mouvements à risques. Enfin, la troisième partie est subdivisée en deux sous-parties pour spécifier les limites et améliorations de la méthodologie et du prototype.

4.2. Préparation de l'espace de données à explorer

4.2.1. Acquisition de bases de données

Une phase d'acquisition des données à explorer est nécessaire dans toute approche de fouille de données. Dans le cadre de l'expérimentation de notre méthodologie, nous avons fait l'acquisition de trois bases de données : une base de données d'accidents maritimes, une base de données météorologiques et une base de données de localisation de navires. Nous présentons ci-dessous ces trois bases de données.

4.2.1.1. Données de Marine Accident Investigation Branch (MAIB)

Nous avons contacté la *Marine Accident Investigation Branch*[79] (MAIB) qui a mis à notre disposition une base de données *Microsoft Office Access* recensant les accidents/incidents de navires qui se sont produits entre 1991 et 2009. Cette base de données est une extraction à partir d'une plus large base de données tenue à jour par le MAIB qui est l'équivalant du bureau sur les événements en mer[80] Français (BEAmer). Ce dernier tient à jour des dossiers papiers décrivant les accidents qui demandent une intégration dans une base de données avant de pouvoir les exploiter.

Les données du MAIB recensent les accidents impliquant les navires britanniques se trouvant n'importe où dans le monde et tous les accidents se trouvant dans les eaux territoriales britanniques. La base de données d'une taille de 16.7 Mo, contient 14 900 cas d'accidents et d'incidents qui concernent 16 230 navires. Dans notre étude nous nous sommes limités aux eaux territoriales britanniques. Le modèle conceptuel de la base de données est représenté sur la Figure 0-1.

[79] http://www.maib.gov.uk/home/index.cfm
[80] http://www.beamer-france.org/index.php

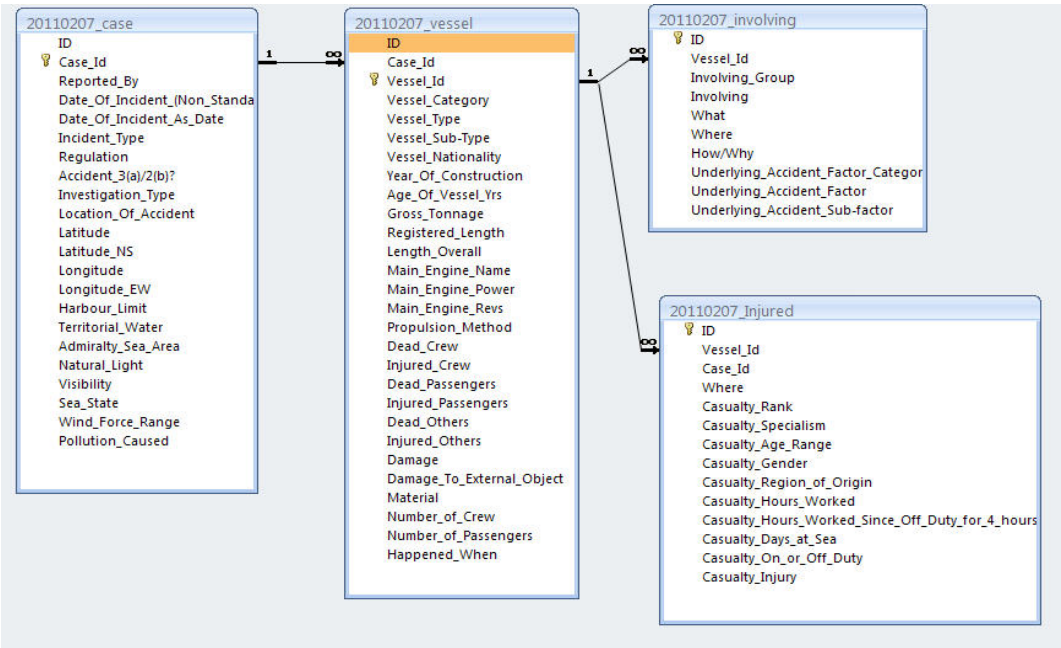


Figure 0-1 : Modèle conceptuel de la base de données MAIB.

4.2.1.2. Données de Modern-Era Retrospective analysis for Research and Applications (MERRA)

Nous avons téléchargé des données météorologiques à partir du site de la National Aeronautics and Space Administration[81]. Cette agence gouvernementale plus connue sous son acronyme NASA, a en charge la plupart des programmes spatiaux civils des États-Unis.

Un historique de données entre 1991 et 2009 a été extrait d'une table appelée IAU 2d atmospheric single-level diagnostics (tavg1_2d_slv_Nx) pour compléter les données météorologiques manquantes de la base de données MAIB. Ces données présentent des mesures de force, direction du vent, pression, humidité, etc. prises toutes les heures du 01/01/1990 au 31/12/2010. La Figure 4-2 illustre l'emprise spatiale des données téléchargées, c'est-à-dire le Royaume-Uni.

Figure 0-2 : Etendue de la zone géographique de téléchargement des données MERRA.

[81] http://disc.sci.gsfc.nasa.gov/daac-bin/FTPSubset.pl?LOOKUPID_List=MAT1NXSLV.

Les données téléchargées concernent les variables suivantes (voir Table 4-1) :

Variable	Description	Unité / Type variable
Lon_Merra	Longitude du point Merra	Degrés décimaux
Lat_Merra	Latitude du point Merra	Degrés décimaux
Merra_time	Timestamp du relevé météo Merra	aaaa-mm-jj hh:mm:ss
SLP	Pression au niveau de la mer	Pa
DISPH	Hauteur de déplacement	m
U2M	Composante Ouest->Est du vent à 2 m au-dessus de la hauteur de déplacement	m.s-1
V2M	Composante Sud->Nord du vent à 2 m au-dessus de la hauteur de déplacement	m.s-1
T2M	Température à 2 m au-dessus de la hauteur de déplacement	Kelvin
QV2M	Humidité spécifique à 2 m au-dessus de la hauteur de déplacement	Kg.kg-1
Winddspeed	Vitesse du vent	m.s-1
Winddir	Direction du vent	Degrés

Table 0-1 : Description des données MERRA téléchargées.

4.2.1.3. Données AIS (Automatic Identification System)

Ce sont des données de cinématiques de navires (Numéro MMSI, vitesse, heure UTC, position, cap, vitesse, etc.) transmises en quasi-temps-réel par les capteurs AIS installés à bord de navires. Ces données transmises sont reçues par d'autres navires navigant à côté et des récepteurs implantés sur les côtes.

Les données AIS anonymes[82], nous sont fournies par DCNS[83] à titre gracieux sous forme de trames d'informations de la National Marine Electronics Association (*NMEA)* envoyées au fur et à mesure de leur réception. Ces données alimente presque en continue notre serveur de données PostgreSQL depuis un an avec un volume de messages d'environ 1.5 Go par jour.

[82] Vu le caractère sensible de l'utilisation de ces données, il a été procédé à leur anonymisation en modifiant les numéros MMSI des navires.
[83] Groupe Français intervenant dans le domaine de l'armement naval et de l'énergie.

4.2.2. Nettoyage des données

De nos jours, d'importantes bases de données sont constituées à partir de flux continus d'informations ou de mises à jour régulières. Ces données peuvent contenir des bruits, des données manquantes, des incohérences ou des imprécisions. En fouille de données, contrairement à l'analyse statistique, les données utilisées sont souvent créées à d'autres fins, elles sont donc souvent inexploitables directement par la fouille de données.

La qualité des connaissances extraites par fouille de données dépend beaucoup de la qualité et de la quantité des données en entrée. En effet, plus il y a de données (cas observés) meilleure est la précision des connaissances. Une analyse statistique des valeurs d'attributs de ces données est nécessaire étant donné que la qualité des résultats d'analyses exploratoires dépend généralement plus de la préparation des données que de la méthode exploratoire utilisée. La première étape de toute investigation dans les données est le calcul des statistiques univariées[84] pour connaître la distribution des variables et identifier les anomalies. Il est important de manipuler les incohérences et les valeurs manquantes avec intelligence car elles peuvent véhiculer des informations intéressantes. En effet, l'absence d'un signal AIS transmis par un chalutier dont la dernière position connue se situe à proximité d'une zone de pêche illégale peut être un indicateur d'une tentative de fraude. Une trajectoire d'un navire aberrante par rapport aux trajectoires du groupe (navires de même type) ou par rapport à ses trajectoires habituelles peut indiquer un comportement suspect.

Avant toute exploration de données, une étape de préparation de ces données est nécessaire pour permettre leur exploitation. Cette préparation est difficile et demande plusieurs itérations compte-tenu de son lien fort avec la qualité des résultats. En effet, la quantité et la qualité des données ont un impact direct et significatif sur la qualité des connaissances générées. Nous nous proposons d'étudier dans les sous-sections suivantes, la distribution des variables pour identifier les anomalies, les corriger et préparer le contexte d'exploration. Les anomalies peuvent être des données manquantes, des incohérences et des imprécisions.

[84] Analyse statistiques concernant une seule variable

4.2.2.1. Données manquantes

Dans le but d'améliorer la qualité des résultats, plusieurs approches peuvent être envisagées pour nettoyer les données comme le remplacement de données manquantes par pondération de moyennes et de médianes ; la prédiction des valeurs manquantes (Jami et al., 2005) et la suppression des observations concernant les valeurs manquantes. Nous avons choisi dans notre cas de remplacer les valeurs manquantes par des valeurs issues d'autres sources de données. Dans le cas de la variable décrivant la force du vent des données d'enquêtes accidents, nous avons remplacé les valeurs manquantes par des valeurs de mesures issues de la base de données MERRA. Comme on le voit sur la Figure 0-3, les coordonnées des données MERRA sont représentées par la grille en bleu et les positions des accidents en orangé. Une requête relie chaque cas de la base de données MAIB avec le point de données MERRA se trouvant à proximité en tenant compte de l'heure la plus proche avec une confiance de 30 minutes. La requête a permis de générer un fichier CSV contenant les résultats de la jointure interne entre les incidents MAIB et les données MERRA.

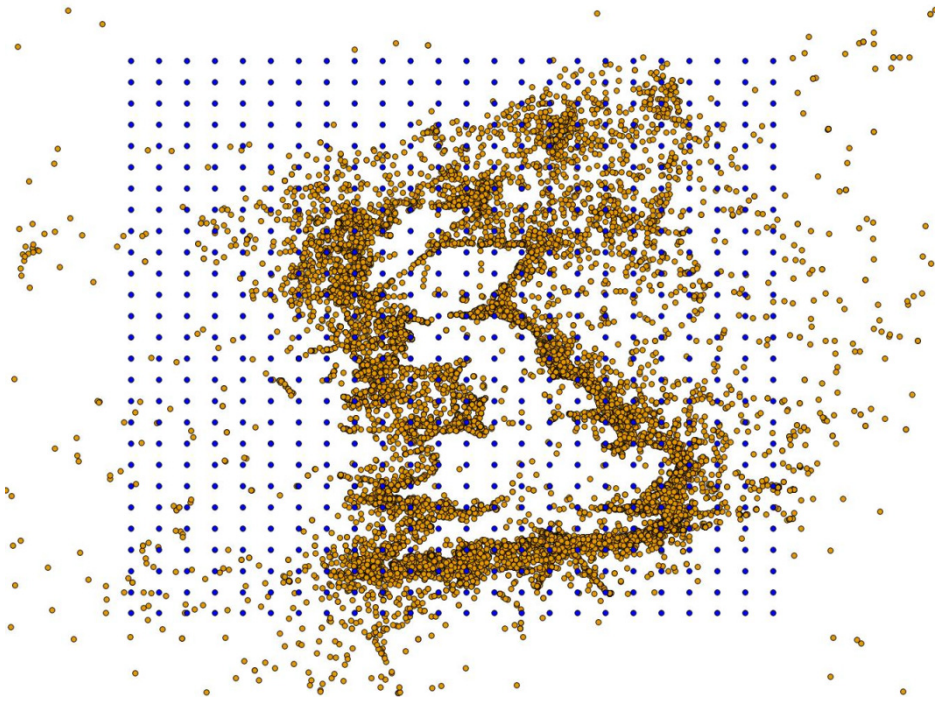


Figure 0-3 : Compléter les données manquantes du MAIB par superposition des données MERRA.

Dans MERRA, la plus petite résolution pour l'intervalle de temps et la région choisis est « 1/2° Latitude x 2/3° Longitude ». Même si cette résolution spatiale semble grande, nous supposons que les mesures météorologiques sont bien interpolées. De plus, sur les 12 000 observations de MAIB bien renseignées, presque 8 000 correspondent à celles de MERRA.

4.2.2.2. Variables continues et distribution hétérogène

Les algorithmes d'extraction de règles d'association ne prennent pas en considération les variables ou les attributs continus dans leur processus d'extraction. Pour ne pas perdre d'informations en entrée de ces algorithmes, une distribution des variables continues a été effectuée pour les séparer en classes d'intervalles. Nous avons choisi les effectifs égaux comme critère de séparation des classes pour éviter de biaiser les résultats de ces algorithmes. En effet, ces algorithmes sont basés sur la découverte d'itemsets fréquents par un calcul d'effectifs, donc une classe ayant plus d'effectifs a plus de chance d'apparaitre dans les règles d'association en sortie.

Nous avons identifié aussi dans la base de données MAIB, des variables discrètes ayant une distribution hétérogène de leur effectif. Cette hétérogénéité va empêcher l'apparition des fameuses pépites d'or ou les cas rares dans les règles d'association en sortie. Les valeurs des variables de la base de données ayant les plus grandes fréquences d'apparition, vont apparaître plus souvent dans les règles au détriment des autres. En prenant l'exemple des navires de pêche dans notre base de données d'enquêtes d'accidents et d'incidents de navires britanniques, nous avions remarqué qu'ils apparaissaient presque systématiquement dans les règles d'association. Ce qui est normal car ils représentent *66%* de l'effectif total de la base de données. Pour faire apparaître les autres catégories de navires, nous avons dû regrouper toutes les catégories de navires en deux grandes classes :

- Classe Transport : Avec un effectif de *34%*, elle regroupe tous les navires de transports de personnes, d'hydrocarbures (Tanker) et de marchandises,

- Classe Pêche : Les navires de pêche représentent un effectif de *66%*.

D'autres algorithmes de fouille de données ne peuvent pas être appliqués sur des variables continues comme le cas de l'algorithme ID3, c'est pour cela que des discrétisations de variables continues sont effectuées.

4.2.2.3. Données aberrantes

Les données aberrantes sont des données atypiques qui sont éloignées de l'ensemble des observations. Elles peuvent être des observations intéressantes pour un domaine et erronées pour un autre. La présence d'une donnée aberrante peut signifier une erreur de saisie, de mesure ou un cas particulier, appelé aussi atypique. L'identification et la décision de garder ou non ces données demande d'avoir une bonne connaissance du domaine d'application étudié.

4.2.2.3.1. Données accidents

La répartition des accidents de la base de données MAIB sur une carte numérique nous a permis par exemple d'identifier et d'écarter les positions aberrantes localisées sur terre, loin des zones de navigation (Figure 0-4). Ces erreurs de positionnement sont peut être dues aux dysfonctionnements de GPS ou à une mauvaise saisie des coordonnées.

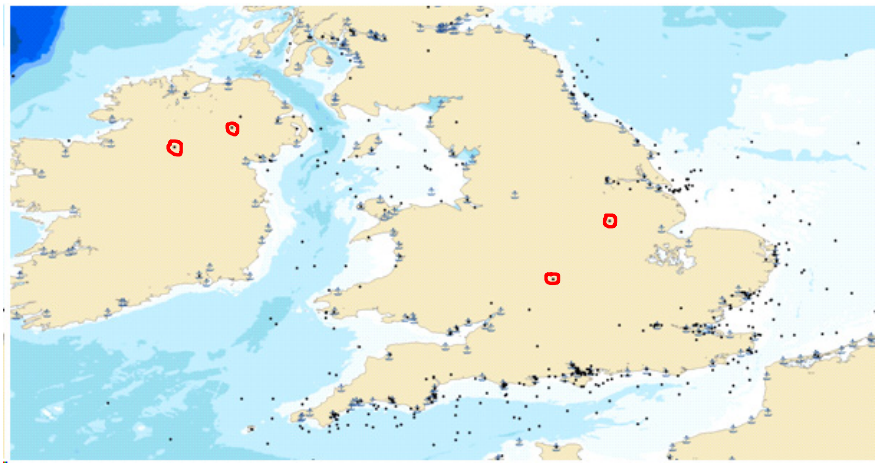


Figure 0-4 : Identification de positions aberrantes par une répartition des positions d'accidents et d'incidents de navires sur une carte numérique. Quelques positions aberrantes sont entourées en rouge.

4.2.2.3.2. Données AIS

Un autre type de valeurs aberrantes a été identifié dans la variable *Age-of-vessel* qui comporte des valeurs négatives. L'analyse des valeurs extrêmes qui représentent les valeurs maximales et minimales de la distribution de la variable, nous a permis de détecter ces valeurs erronées pour les éliminer de la base.

Nous avons suivi la même approche que la répartition des accidents MAIB sur une carte numérique pour détecter les trajectoires aberrantes. La reconstitution des données AIS sous forme de trajectoires sur une cartographie, nous a permis de découvrir des trajectoires aberrantes qui passent sur les terres.

La Figure 4-5 montre l'exemple d'une trajectoire aberrante passant entre le port de Malaga et Valence. Cela est dû la plupart du temps à une perte de signal AIS entre deux points éloignés.

Figure 0-5 : Une trajectoire aberrante due à une perte de signal AIS.

Pour nettoyer ces données, nous avons choisi une approche simple. On affiche sur une cartographie toutes les trajectoires d'un jeu de données puis les trajectoires passant sur terre sont sélectionnées pour récupérer leur identifiant MMSI. Les trajectoires ayant des partitions aberrantes sont alors supprimées à partir du fichier de données source. Cette procédure est simple et rapide pour un nombre raisonnable de trajectoires. Pour les trajectoires passant sur les terres à proximité de la mer, une interpolation est effectuée pour compléter ces trajectoires. Après suppression des trajectoires aberrantes, interpolation des trajectoires qui passent près de la mer dans certaines partitions et découpages des trajectoires aux ports, nous avons conservé 65 trajectoires de tankers.

Il y a d'autres données aberrantes dues à la perte de signaux. Ces trajectoires ne passent pas par les terres mais présentent des partitions très longues. L'algorithme de détection de trajectoires aberrantes va ainsi détecter ces partitions alors qu'elles ne nous intéressent pas. Ces aberrations sont dues à une perte de signale et non aux déplacements réels des navires. Nous proposons pour éliminer ces aberrations, une limitation de la

longueur des partitions. Des conditions sont fixées dans le calcul des partitions aberrantes pour filtrer les parties jugées trop longues ou trop courtes. Pour éviter des analyses statistiques couteuses, nous avons fixé des seuils arbitrairement : si la partition est plus petite que 0.00009° - environ 10 mètres - ou plus grande que 0.09° - 10 kilomètres environ - elle n'est pas prise en compte dans le calcul.

4.2.3. Modélisation des espaces de données

Nous présentons ci-après la modélisation des espaces de données concernant les données accidents et les données AIS.

4.2.3.1. Données Accidents

Nous avons sélectionné dans la base de données MAIB, les données qui décrivent les accidents (type d'accidents, position, temps, eaux territoriales, etc.), les caractéristiques des navires (identifiant IMO, type du navire, âge du navire, longueur, etc.) et la description de l'environnement (visibilité, état de la mer, force du vent, etc.). La sélection des variables sur lesquelles va porter notre analyse va réduire le nombre de variables à considérer, le nombre de règles générées et ainsi faciliter l'interprétation des résultats. Cette sélection de données va constituer le contexte d'exploration sur lequel va porter l'extraction de règles d'association dans le but de trouver les relations d'implications entre les différents facteurs de situations.

Quelques variables prises toutes seules permettent déjà d'évaluer le risque comme le type du navire. En effet, les autorités maritimes s'appuient beaucoup sur cette dimension pour mesurer le risque. L'entrée d'un navire militaire et d'un cargo dans la rade de Brest par exemple, n'est pas perçue au même niveau de risque par les autorités maritimes. On imagine alors l'apport d'une mise en relation entre plusieurs variables à l'évaluation des risques.

Dans l'analyse des accidents de navires, nous avons gardé les types d'accidents les plus fréquents et pouvant avoir une relation avec les conditions météorologiques, à savoir, l'échouement et le naufrage.

En ce qui concerne les variables, nous avons choisi de garder les conditions de navigation (vent, courant, etc.), les caractéristiques de navires (type, âge, etc.) et les risques maritimes (type du risque, catégorie, etc.) pour la détection de facteurs de risques et la localisation (latitude et longitude) de l'accident ainsi que le type d'accident pour la détection des zones de risques.

Nous avons aussi enlevé les données fluviales de la base de données MAIB car nous travaillons sur une problématique maritime.

Les données résultant de la sélection effectuée sur la base de données MAIB contiennent les attributs suivants (Table 4-2) :

Attribut	Description	Unité / Type variable
Case_id	Identifiant de l'incident	Entier
Incident_type	Type de l'incident	Texte
Vessel_id	Identifiant du navire impliqué	Entier
Vessel_Category	Catégorie de navire impliqué	Texte
Age_Slice_Of_Vessel	Intervalle d'âge du navire impliqué	Text
Incident_time	Timestamp de l'incident	aaaa-mm-jj hh:mm:ss
Location	Localisation de l'accident (Coastal waters, High seas, Non-tidal waters, Port/harbour area)	Texte
Territorial_Water	Indique l'eau territoriale de l'accident	Texte
Lat_vf	Latitude de l'incident	Degrés décimaux
Lon_vf	Longitude de l'incident	Degrés décimaux
Sea_state	Etat de la mer	Texte
Wind_force	Force du vent	Echelle Beaufort
Visibility	Visibilité au moment de l'accident (Poor, Good, Mod.)	Texte

Table 0-2 : Description des attributs de la base de données MAIB sélectionnés et préparés.

4.2.3.2. **Données AIS**

Concernant les données de déplacement, nous avons dû créer plusieurs fichiers de données selon l'algorithme d'exploration car chacun possède un format de données d'entrée particulier. Les données ont été aussi limitées par type de navires car les

comportements et plus particulièrement les mouvements sont différents d'un type de navire à un autre. Les tankers par exemple ont tendance à faire des trajets presque rectilignes entre deux ports (chemin le plus court) alors que les navires de pêche font des déplacements souvent spécifiques autour d'un port.

La préparation des données est certes une condition nécessaire mais elle n'est pas pour autant suffisante pour avoir des connaissances de qualités. En effet, d'autres conditions doivent être satisfaites comme par exemple le choix d'une méthode d'exploration pertinente.

Nous présentons ci-après, un tableau présentant quelques attributs de la base de données AIS (Table 4-3) :

Attributs	Description
MMSI	Identifiant unique du navire
IMO	Identifiant de l'équipement AIS
Nom	Nom du navire
Type	Type du navire
Latitude	Latitude du navire au moment de l'envoi du signal AIS
Longitude	Longitude du navire au moment de l'envoi du signal AIS
Vitesse	Vitesse du navire au moment de l'envoi du signal AIS
Cap	Direction du navire au moment de l'envoi du signal AIS
Horodatage	Timestamp de l'envoi du signal AIS
Cargaison	Type de cargaison transportée par le navire

Table 0-3 : Description de quelques attributs des données AIS.

4.3. Extraction de comportements à risques

L'extraction de comportements à risques passe préalablement par l'extraction et l'interprétation des situations et mouvements de navires. Dans les sous-sections suivantes, nous allons commencer par l'extraction de quelques exemples de situations à risques et mouvements à risques à partir de données réelles d'enquêtes d'accidents maritimes et de déplacements de navires.

4.3.1. Extraction de situations à risques

Parmi les méthodes d'extraction de situations à risques que nous avons choisies dans le chapitre précédent (Cf. section 3.2.3.1) pour extraire des connaissances sur les facteurs et les zones à risques, certaines ne sont pas intégrées à ShipMine. C'est le cas des méthodes d'extraction de règles d'association et la construction d'arbres de décision. Ces méthodes sont utilisées à partir de programmes tiers. La méthode de découverte de zones à risques quant à elle est intégrée à ShipMine.

L'exploration des données d'enquêtes d'accidents maritimes de MAIB, peut permettre la génération de connaissances sur des situations à risques mettant en relation des conditions de navigation (vent, courant, etc.), des caractéristiques de navires (type, âge, etc.) et les risques maritimes (type du risque, catégorie, etc.). La découverte de ces relations peut aider à la prédiction et au ciblage des accidents maritimes. Cette exploration peut aussi permettre la découverte de connaissances sur les zones à forte densité d'accidents pour l'identification de zones à risques.

La discussion des résultats découverts automatiquement par ces trois méthodes d'extraction de situations est présentée ci-après.

4.3.1.1. Extraction de règles d'association

La base de données de transactions utilisée est la base de données des accidents et incidents maritimes du MAIB (Cf. section 4.2.3). Les *items* sont les différents facteurs comme le type du navire, la force du vent, la localisation relative et les types d'accidents. La découverte des associations dans cette base consiste à chercher les ensembles *d'items* fréquemment liés dans les accidents (Idiri & Napoli 2012a).

L'exploration des données MAIB par l'algorithme *Apriori* après les avoir préparées (Cf. section 4.2.3), a permis de découvrir beaucoup de règles. Ces règles ont été obtenues en faisant varier les valeurs des seuils du support "*minsupp*" (indicateur de fiabilité de la règle) et de la confiance "*minconf*" (indicateur de précision de la règle). L'exploration des 4 247 observations a permis de générer :

- 631 règles avec des seuils de support et de confiance supp=0.04 et conf= 0.6,

- 371 règles avec supp=0.1 et conf= 0.6,

- 803 règles sur une sélection de 715 observations concernant les navires de transport avec supp=0.05 et conf= 0.5.

La sélection des règles intéressantes est effectuée en deux étapes. La première étape est une sélection selon les mesures de support, de confiance et de lift. La deuxième étape est effectuée sur un critère de pertinence par rapport à l'analyse des situations à risque. Dans la première étape, nous avons sélectionné plusieurs règles intéressantes en termes de métrique (support, confiance et lift) mais il s'est avéré dans la deuxième étape, après une analyse de pertinence que la majorité de ces règles n'apportaient pas de nouvelles connaissances.

Nous présentons ci-dessous, une règle d'associations par classe de résultats obtenus :

- **Règle 1** (Règle de prédiction) : Location = Coastal waters, Vessel_Category = Fish catching/processing, Age_Slice_Of_Vessel = 11 to 18 years → Incident_Type=Machinery Failure (supp=0.086 ; conf= 0.725 ; lift=1.47),

  La première règle informe que les accidents de navires de pêche âgés de 11 à 18 ans et navigant dans les eaux côtières britanniques sont causés dans 72% des cas par une panne mécanique. Cette règle est assez fréquente, elle représente presque 9% de la base de données MAIB.

- **Règle 2** (Règle de ciblage) : Vessel-Category=Fish catching → Vessel-Type= Trawler (supp= 0.14 ; conf= 0.43 ; Lift= 3),

  La deuxième règle informe que si un accident concerne un navire de pêche alors dans 43% des cas c'est un chalutier. La fiabilité de cette règle est vérifiée par 14% des cas de la base de données. Selon un sous-officier de la marine marchande, les chalutiers sont les plus exposés au risque de naufrage car ils tirent un chalut qui peut s'accrocher et entrainer vers le fond le chalutier. Donc cette règle confirme une information connue auparavant par les navigateurs.

- **Règle 3** (Règle Banale) : Vessel-Category=Passenger → Pollution-Caused=No (supp=0.15 ; conf= 0.73 ; lift= 1.2)

  La dernière règle, présente une règle triviale (inutile) qui signifie que les accidents de navires transportant des voyageurs ne causent pas de pollution dans 73% des cas, ce qui semble logique car ils transportent des passagers et non des substances polluantes. Il est important de noter que la notion de pollution ou non est ici relative aux autorités maritimes britanniques qui décident ou non si une pollution est avérée.

Concernant les règles impliquant des conditions météorologiques, océanographiques et de contexte, la majorité des règles trouvées, ayant le support (Supp >= « minsupp ») et la confiance (Conf >= « minconf ») mettent en relation des conditions normales : mer calme, force de vent faible, dans les ports, etc. voici ci-après, un exemple de quelques règles de ce type (Table 0-4).

Partie 1	Partie 2	Supp	Conf
Location=Port/harbour area, Wind_Force_Range=0-3 (calm)	Vessel_Category=Transport	5.7%	70%
Incident_Type=Grounding, Sea_State=Calm <2 ft	Vessel_Category=Fish catching/processing	4.6%	66%
Sea_State=Sheltered Waters, Vessel_Category=Transport	Location=Port/harbour area	8%	57%

Table 0-4 : Exemple de règles d'associations impliquant des conditions météorologiques, océanographiques et de contexte normales dans les accidents.

L'un des inconvénients de cette exploration est le fait que les *items* sont tous au même niveau, c'est pour cela que toutes les implications possibles ayant le seuil de support et de confiance sont générées. On se retrouve avec énormément de règles qu'il faut analyser pour identifier celles qui décrivent de nouvelles connaissances auxquelles les experts n'avaient pas pensés auparavant, des banalités (par exemple, il fait froid en hiver et chaud en été) et des idées reçues.

Par la suite, nous allons utiliser les arbres de décision pour définir une variable à expliquer à partir d'autres variables dites explicatives. Cela va focaliser la recherche sur

les règles permettant d'expliquer les types d'accidents par rapport aux facteurs (variables) d'environnement, de contexte et de caractéristiques des navires.

4.3.1.2. Construction d'un arbre de décision

Nous présentons ci-dessous le résultat obtenu de l'exploration des données MAIB par les arbres de décision. Comme on le voit sur la Table 4-5 la représentation graphique de l'ensemble des règles obtenues permet déjà une exploitation plus aisée et rapide des règles d'implication.

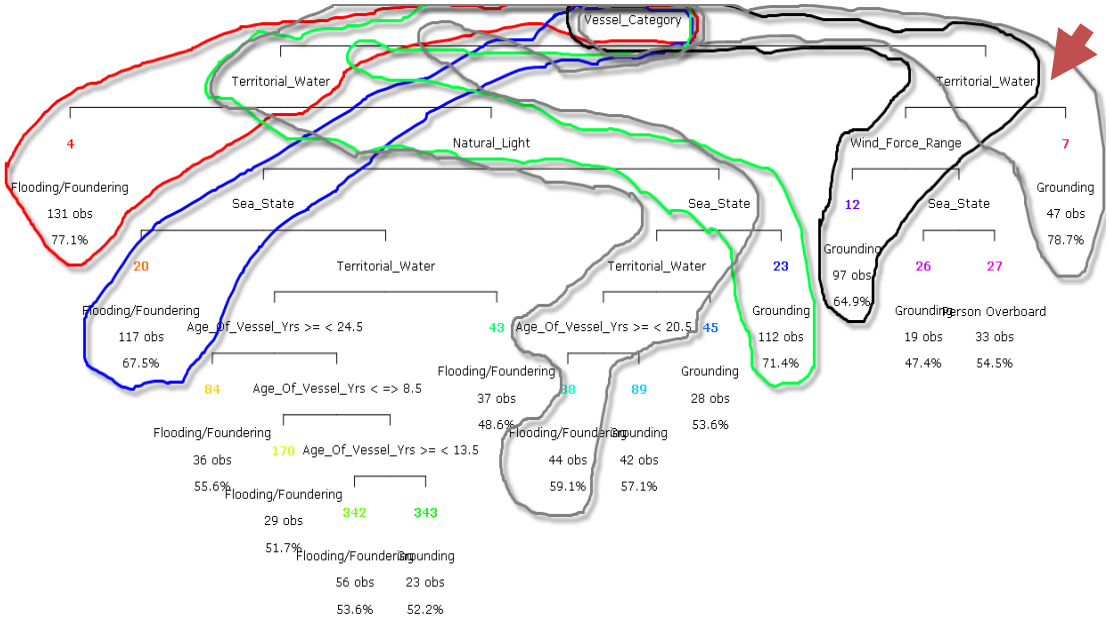


Figure 0-6 : Arbre de décision expliquant les types d'accidents par rapport à des facteurs météorologiques, océanographiques et des caractéristiques de navires.

Prenons par exemple, la règle 7 qui est entourée en gris sur l'arbre de décision (Figure 4-6). Cette règle identifie une relation entre les échouements de navires de transport britanniques et des localisations relatives : les accidents de navires de transport se trouvant sur les territoires : Northen Irish, Scottish ou Welsh sont des échouements dans 78,7% des cas.

L'utilisation des arbres de décision n'a pas été plus fructueuse que celle des règles d'association. La majorité des règles trouvées montre que les accidents ont eu lieu dans

des conditions météorologiques, océanographiques et de contextes les plus normales possible : mer calme, force de vent faible, bonne visibilité.

Pour confirmer que ces résultats ne sont pas dus à une sous-estimation des conditions météorologiques au cours de l'enquête accident et incident, nous les avons comparées aux données MERRA. Après vérification, il s'est avéré qu'il y a eu plutôt une surestimation des données MAIB, comme il est possible de le percevoir sur le tableau croisé des 4 334 observations différentes entre les données MAIB et MERRA (Table 4-5).

Table de wind_force par wind_force_Merra

		wind_force_Merra(wind_force_Merra)			
		0-3	4-6	7-9	Total
wind_force					
0-3	Pctage en ligne	0.00	99.73	0.27	
10-12	Pctage en ligne	8.61	60.93	30.46	
4-6	Pctage en ligne	99.60	0.00	0.40	
7-9	Pctage en ligne	7.07	92.93	0.00	
Other	Pctage en ligne	56.70	41.24	2.06	
Total	Fréquence	1915	2361	58	4334

Généré par le Système SAS ('Local', XP_PRO) le 07 février 2013 à 3:51:25 PM

Table 0-5 : Comparaison des forces du vent mal renseignées pendant l'enquête accident avec ceux de la base de données MERRA.

Le problème vient probablement de l'hypothèse de départ qui suppose une relation entre les mauvaises conditions météorologiques et les accidents maritimes. Il est possible que cette hypothèse ne soit pas si souvent vérifiée, c'est pour cela que les résultats ne sont pas ceux attendus.

Selon les résultats, il est possible que les accidents maritimes aient une relation avec la localisation. L'analyse des localisations d'accidents et incidents maritime va être abordée dans la section suivante.

4.3.1.3. Extraction de zones à risques

Des zones denses ont été découvertes par regroupement des localisations des accidents et incidents maritimes. Ces zones peuvent être utilisées pour suivre l'évolution des risques, suivre de plus près les navires qui fréquentent ces zones (Vessel Of Interest) et avoir une meilleur planification des moyens maritimes de surveillance et d'intervention.

L'utilisation de la fonctionnalité « Zones à risque » intégrée à ShipMine sur les données de localisation des accidents et incidents du MAIB, nous a permis d'identifier des zones accidentogènes. Pour une distance de voisinage égale à 1km et un seuil minimum d'accidents égal à 50, nous avons obtenu trois zones à risques localisées respectivement à côté des villes suivantes : Portsmounth, Milford Haven et Bournemouth. Comme nous le voyons sur la Figure 4-7, la zone autour de Portsmounth contient 178 accidents, Milford Haven 102 accidents et Bournemouth 61 accidents survenus entre 1991 et 2009.

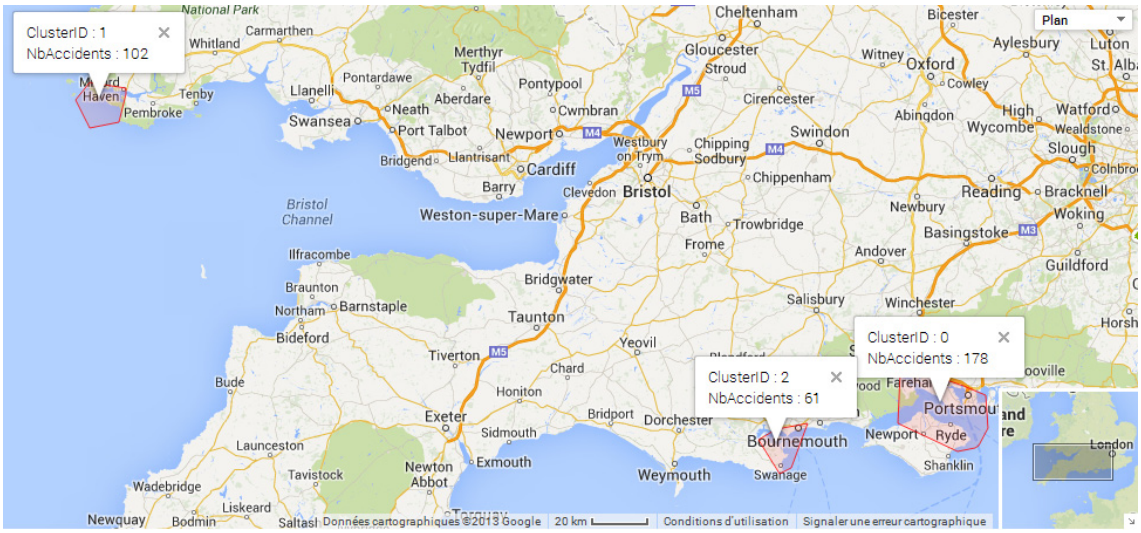


Figure 0-7 : Découverte de zones accidentogènes dans le sud de l'Angleterre pour une distance de voisinage de 10 km et un minimum de 50 accidents.

Plus nous augmentons la distance de voisinage, plus les zones accidentogènes deviennent larges et leur nombre augmente ou diminue selon la distribution des accidents. Comme on le voit sur la Figure 4-8-(a), pour une distance égale à 20 km nous avons obtenu 9 zones. La zone de Portsmounth vu dans le premier exemple est passée à 235 accidents et le nombre de zones a augmenté car avec une distance de 20 km, il y a plus de concentrations d'accidents identifiées.

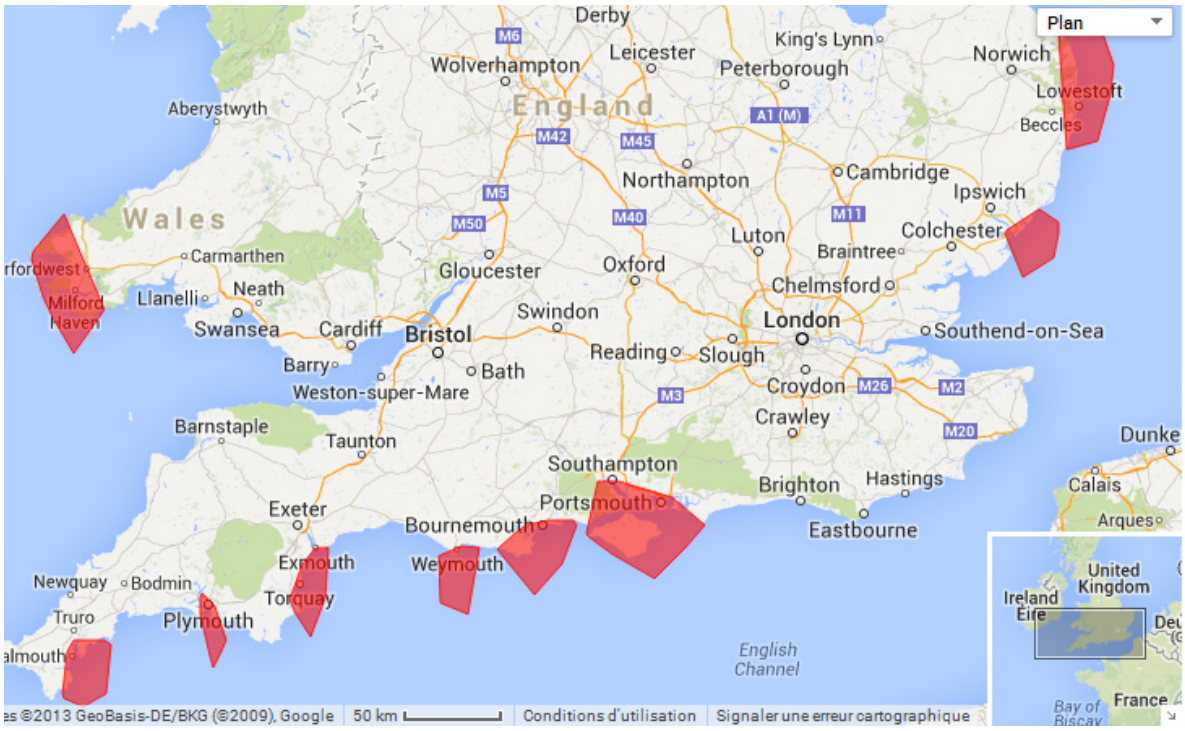


Figure 0-8 : Evolution de la largeur des zones accidentogènes par rapport à la distance de voisinage D=20 km.

Il peut s'avérer judicieux d'afficher les zones à risques sur des ENC pour identifier d'éventuelles relations entre des objets maritimes et ces concentrations d'accidents, comme par exemple des rochers affleurants.

4.3.2. Extraction de mouvements à risques

Dans cette section, nous allons procéder au test d'extraction de mouvements potentiellement à risques et plus particulièrement à la découverte de motifs de trajectoires aberrantes, de navigations proches et de motifs de routes de navigation à partir des historiques de déplacements de navires (position, vitesse, cap, etc.). La procédure de test utilisée est exposée ci-après.

4.3.2.1. Trajectoires aberrantes

L'exploration des 65 trajectoires de 16 tankers se déplaçant en Méditerranée en utilisant la fonctionnalité de ShipMine permettant de détecter les trajectoires anormales a permis d'extraire des mouvements anormaux de navires.

Pour une distance de voisinage égale à 100 mètres et une proportion de trajectoires non ressemblantes dans le voisinage égale à 98%, nous avons obtenu presque 5 000 partitions anormales. La Figure 4-9 présente un focus sur le détroit de Gibraltar[85] qui présente quelques comportements anormaux. Nous distinguons des changements de cap brusques, des attentes loin des ports et des changements de destinations à quelques encablures du port. Ces comportements sont anormaux et peuvent décrire un risque. Nous allons discuter par la suite quelques exemples de motifs de mouvements que nous avons obtenus.

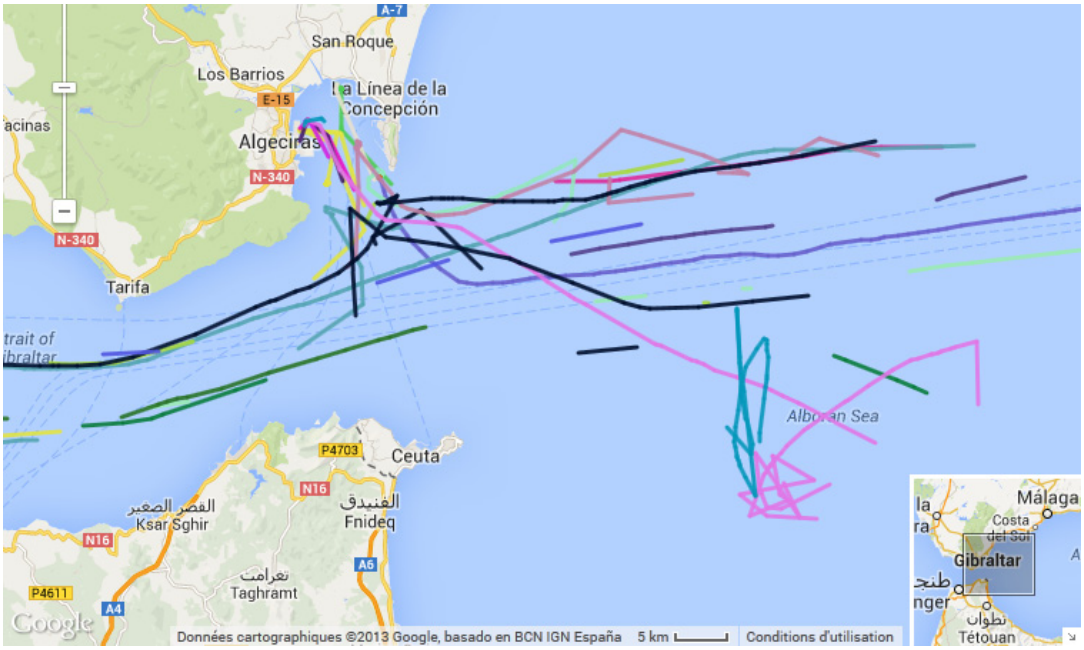


Figure 0-9 : Détection de comportements anormaux de tankers à Gibraltar.

Prenons le cas d'un tanker qui a l'air d'attendre à l'écart du passage de Gibraltar. Ce navire a eu deux comportements consécutifs similaires comme on peut le voir sur la Figure 4-10. Que signifie ce comportement ?

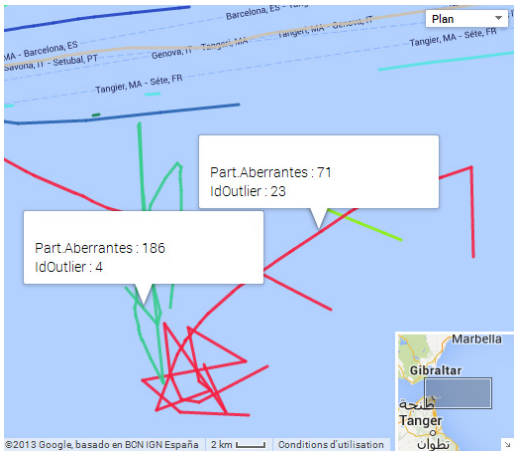


Figure 0-10 : Comportement anormal d'un navire

Ce comportement peut être juste une attente pour rentrer au port à défaut de place

[85] De l'arabe Jabel Tariq, c'est un passage maritime situé au sud de l'Espagne et qui relie entre la mer Méditerranée et l'océan Atlantique.

ou d'autorisation. Il peut être aussi en négociation pour vendre la marchandise transportée au plus offrant comme il peut être un comportement à risque (avarie, trafic illégal, etc.).

Le fait que ce tanker reste loin des ports habituellement fréquentés par ce type de navires, lève des doutes sur ses objectifs réels. En effet, ce comportement peut représenter un risque d'échange frauduleux entre navires ou à partir de la côte en utilisant des embarcations rapides. Le tanker est à 25 km des côtes. De plus, un comportement peut présenter des risques sur le trafic maritime surtout la nuit où il n'y a pas de visibilité.

Le même comportement près des ports (Figure 4-11), peut être considéré comme une attente de déchargement d'une partie ou de toute la marchandise vers un navire plus petit. Ce genre de procédure se fait dans les ports qui ne sont pas adaptés pour recevoir de grands navires. Justement, les Figures suivante (Figure 4-12 et Figure 4-13) montrent des mises en couple[86] réelles entre deux tankers dans un port de Gibraltar.

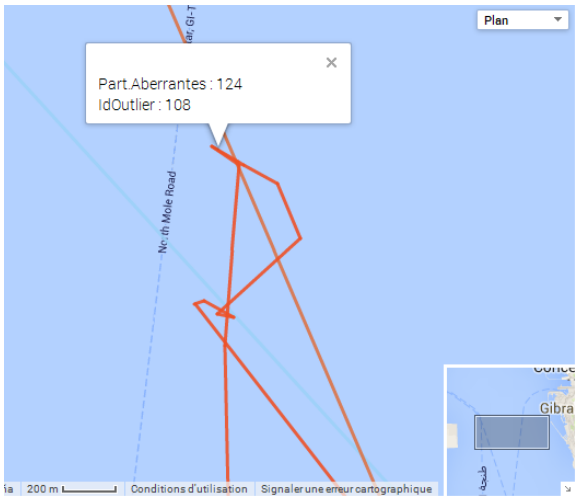


Figure 0-11 : Comportement d'attente et de mise en couple d'un tanker pour déchargement de marchandise.

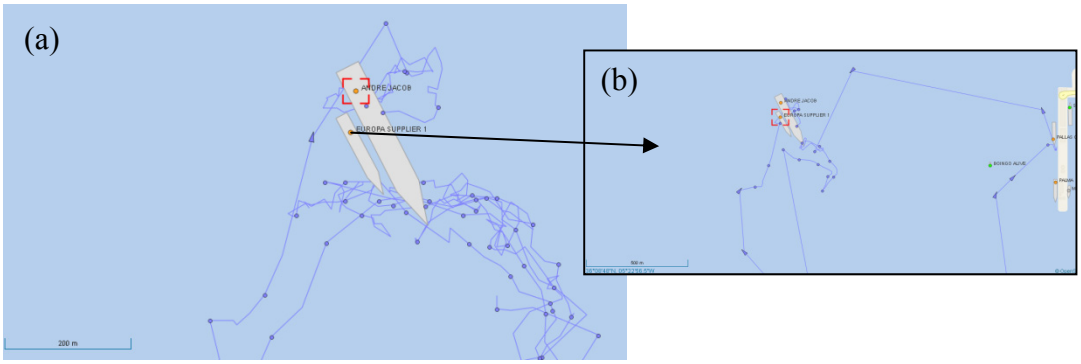


Figure 0-12 : Comportement de déchargement de marchandise -Le grand tanker se met en couple (a) avec un navire qui vient du port (b). (http://www.vesselfinder.com/fr)

[86] Expression utilisée par les navigateurs pour designer deux navires se mettant l'un à côté de l'autre.

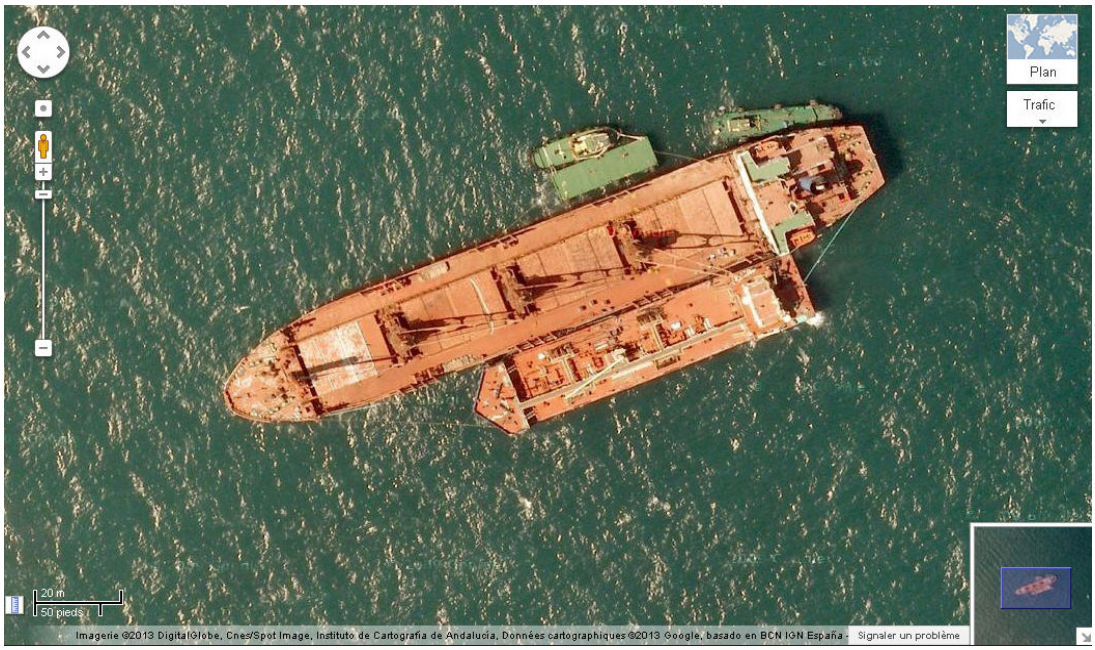


Figure 0-13 : Mise en couple de deux tankers dans un port de Gibraltar
(source Google Maps)

Nous présentons un autre comportement anormal sur la Figure 4-14. Ce comportement peut indiquer une manœuvre d'un tanker pour récupérer un objet tombé à la mer ou pour prendre de la distance avec un autre navire. Ce genre de manœuvre est anormale, risquée et peut engendrer un accident du trafic.

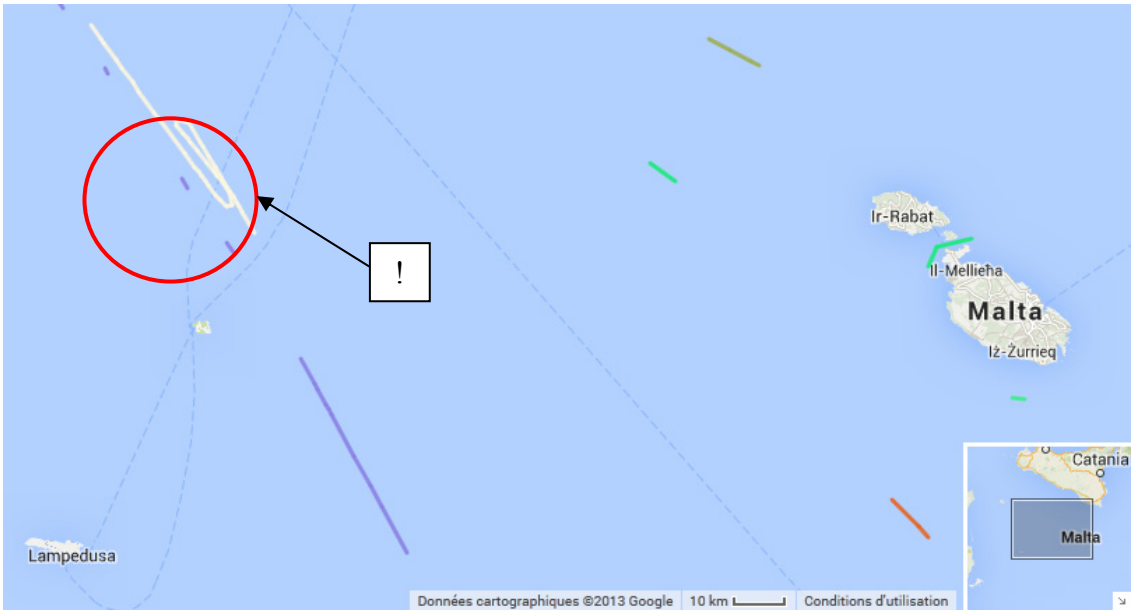


Figure 0-14 : Comportement d'un navire qui change plusieurs fois de destinations.

La Figure 4-15 montre un autre comportement anormal d'un tanker qui sort du port puis qui y revient après plusieurs kilomètres de navigation. Ce comportement peut représenter l'état d'un navire qui a nécessité un retour d'urgence au port pour cause d'avarie par exemple. Nous présentons ci-dessous (Figure 4-16) une sous-trajectoire aberrante qui présente un cas ressemblant à la dérive d'un tanker. Le navire a une trajectoire inhabituelle qui tend vers les côtes puis elle s'arrête à 6 km de la côte.

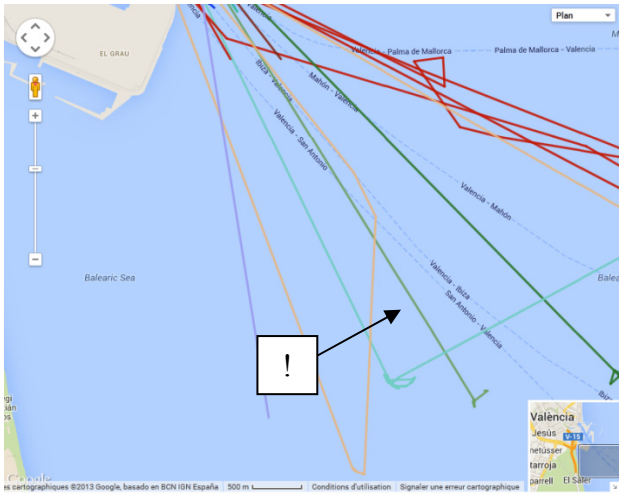


Figure 0-15 : Comportement d'un tanker qui revient au port.

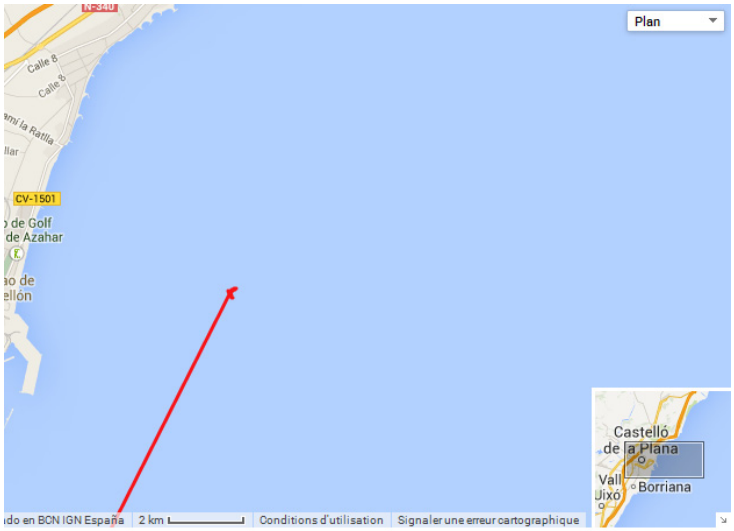


Figure 0-16 : Comportement d'un tanker qui ressemble à une dérive.

Nous avons vu au travers de quelques exemples présentés dans cette section, que la méthode de découverte de trajectoires et sous-trajectoires aberrantes utilisée permet bien d'extraire des motifs de mouvements anormaux qui peuvent décrire des comportements à risques.

4.3.2.2. Navigation proche

L'exploration d'un historique de déplacement de 9 navires de pêches navigants dans les eaux territoriales des îles Féroé (Figure 4-17) a permis d'identifier des comportements de navigations parallèles proches. Ces comportements peuvent indiquer des pêches parallèles qui sont interdites.

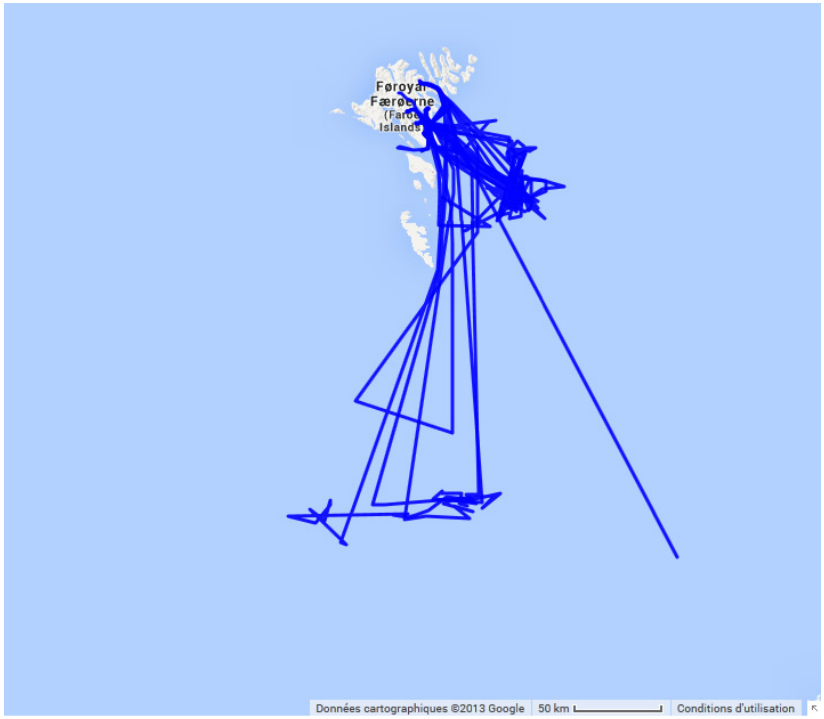


Figure 0-17 : Historique de trajectoires de navires de pêche navigant dans les eaux territoriales des îles Féroé.

L'exploration de ces déplacements de navires de pêche en utilisant la fonctionnalité « Navigation proche » de ShipMine avec un nombre de trajectoires minimal égal à 2, une distance de voisinage maximale égale à 1 kilomètre et une durée minimale de navigation proche égale à 10 minutes, nous a permis de découvrir 8 navigations parallèles. La simplification des trajectoires par l'algorithme Douglas-Peucker temporel (Cf. section 2.2.3.1.4 du chapitre 2) a été effectuée avec une précision (tolérance) de 40 mètres. Nous avons choisi cette valeur par visualisation de différents résultats de trajectoires simplifiés obtenus pour plusieurs valeurs de tolérance. L'idée est de simplifier les trajectoires mais sans les rendre trop lisses.

Nous présentons sur la figure suivante (Figure 4-18), deux comportements de navigation proches pouvant décrire une pêche parallèle. Dans la navigation parallèle numéro 7 par exemple, les deux navires sont restés en parallèle plus de 30 minutes à la date du 15 juin 2013.

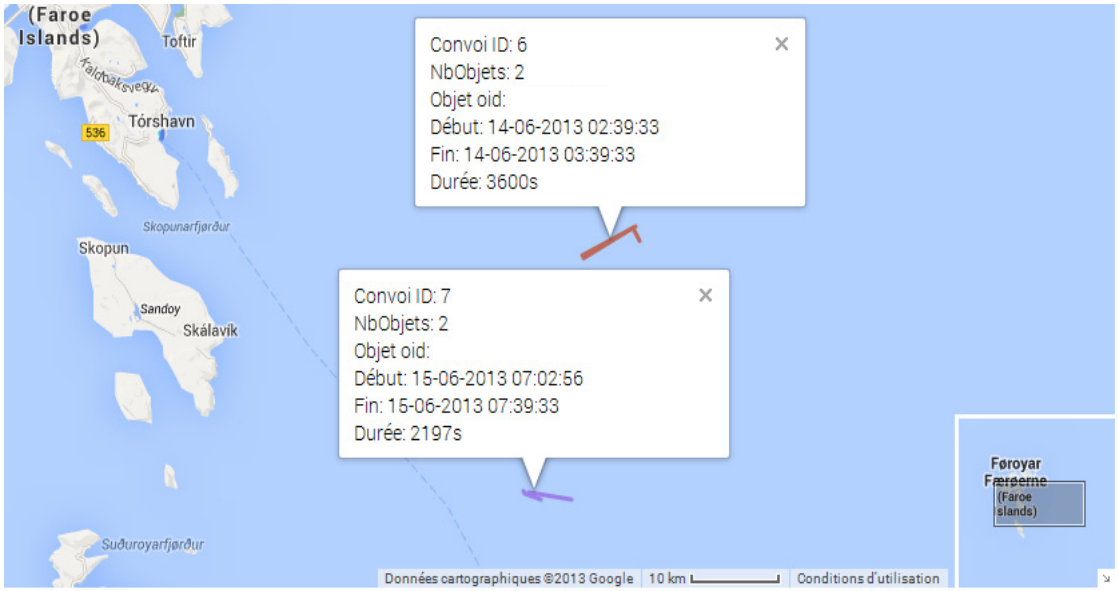


Figure 0-18 : Découverte de pêches parallèles de navires navigant dans les eaux territoriales des îles Féroé.

4.3.2.3. Routes de navigation maritime

Dans cette section, nous allons extraire des trajectoires de navigation habituelles ou type à partir de l'exploration de 14 trajectoires de 2 navires de pêche ayant navigué à proximité du port de Sète. Les traces de ces navigations sont présentées sur la Figure 4-19.

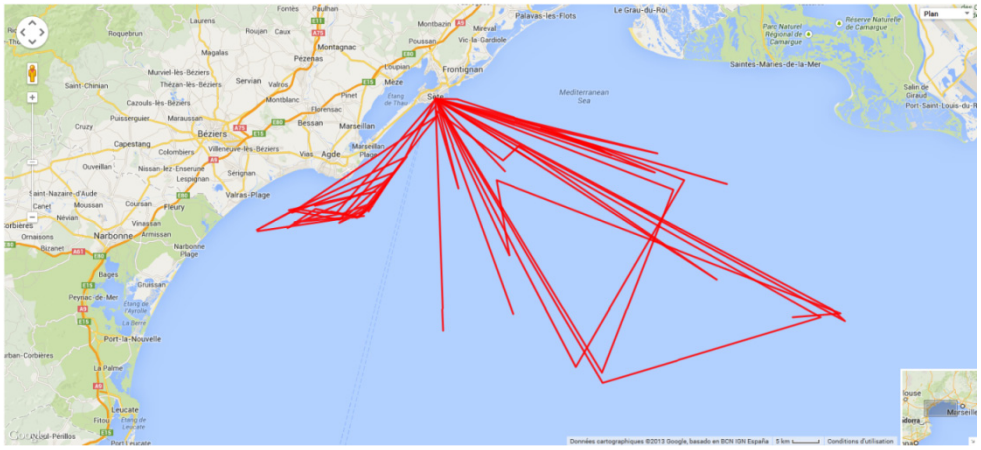


Figure 0-19 : Historique de traces de navigation de 2 navires de pêche dans le port de Sète.

L'exécution de la fonctionnalité « Routes de navigation » de ShipMine sur les 14 trajectoires de navires avec une distance de voisinage de 15 km et un minimum de partitions voisines MinLns égal à 2, nous a permis de découvrir une agrégation des déplacements de ces navires. La route de navigation découverte agrège le comportement de 12 trajectoires comme on peut le percevoir sur la Figure 4-20. Cette route de navigation décrit le comportement général des navires de pêche analysés. Ces navires ont l'habitude de faire des allers-retours entre le port de Sète et la zone sud-est ou sud-ouest avec une navigation parallèle au cap d'Agde. Etant donné que les navires de pêche partent du port de Sète, les deux trajectoires type ont été connectées par densité.

Figure 0-20 : Découverte de routes de navigation de navires de pêche naviguant à proximité du port de Sète.

Les motifs spatiaux décrivant les trajectoires type de navigation peuvent être utilisés pour découvrir les trajectoires aberrantes qui s'écartent du comportement normal. Ces trajectoires anormales peuvent décrire un comportement à risque (cf. section 4.3.2.1) comme ils peuvent décrire des routes à risque de naufrage si on se focalise sur les chalutiers ou des routes à risque de pollution si on se focalise sur les navires de transport de matières dangereuses.

4.4. Limites et pistes d'amélioration

Nous allons exposer dans la section suivante les limites liées à notre proposition méthodologique, à l'atelier mis en œuvre ainsi que quelques pistes d'amélioration possibles.

4.4.1. Méthodologie

La méthodologie utilisée est basée sur la fouille de données qui ne fait pas participer les utilisateurs dans l'exploration de données. C'est une sorte de boite noire pour les utilisateurs. Nous présentons dans les perspectives, une proposition d'amélioration de notre méthode.

La validation de notre méthodologie d'aide à l'extraction de connaissances sur les comportements à risques pourrait être faite sur le terrain avec des experts maritimes comme les capitaines de navire[87]. Cela permettrait de savoir si les connaissances découvertes automatiquement correspondent bien aux idées reçues et aux intuitions des navigateurs. La coïncidence de connaissances permettrait de réconforter les connaissances des navigateurs et valider notre méthodologie. Les connaissances qui ne coïncideraient pas, pourraient être analysées pour en comprendre les raisons : est-ce que c'est dû au fait que les connaissances générées mettent en évidence des situations et mouvements à risques nouveaux, auxquels les navigateurs n'ont pas été confrontés auparavant ou sont-elles biaisées ? Dans ce dernier cas, il serait intéressant d'identifier les causes qui ont amené à générer de telles connaissances.

Les règles d'association n'ont pas permis la découverte de relations négatives entre les itemsets du genre « si A alors non B » (trouver A exclut B et inversement). Les

[87] Contacter par exemple l'Association Française des Capitaines de Navires (http://www.afcan.org/presentation1.html).

relations d'implications des règles d'association sont monodirectionnelles et ne permettent pas la découverte de ce genre de règles. L'extraction de relations de corrélations (Silverstein et al., 1998) peut permettre de découvrir des relations négatives, positives et nulles entre les itemsets. Il serait intéressant d'utiliser cette méthode pour la suite de nos travaux.

Les méthodes de groupement par densité utilisées ne prennent pas en compte les obstacles. Les zones à risques extraites, couvrent des territoires non maritimes. A cette valeur de distance, toutes les localisations d'accidents sont accessibles par densité. Une amélioration possible est d'utiliser des méthodes de clustering spatial prenant en compte cette contrainte d'existence d'obstacles (Tung et al., 2001).

4.4.2. ShipMine

L'atelier ShipMine intègre aujourd'hui quatre algorithmes qui sont DBSCAN, TRAOD, TRACLUS et Convoy. Ces algorithmes supportent des fonctionnalités d'extraction de zones à risques, d'identification de trajectoires aberrantes, de routes de navigation et de navigations proches qui ont permis d'obtenir des résultats prometteurs sur nos données maritimes (cf. section 4.2). Dans une perspective d'amélioration des fonctionnalités de ShipMine, d'autres algorithmes ont été sélectionnés pour leur éventuelle intégration dont deux testés sur les données MAIB, à savoir Apriori et ID3 (cf. sections 4.3.1.1 et 4.3.1.2).

Dans ce travail, nous ne nous sommes pas attardés sur l'étude de la scalabilité, l'optimisation des algorithmes et des programmes de ShipMine pour une utilisation dans le monde opérationnel. Les performances de ShipMine sont suffisantes pour nos besoins expérimentaux mais doivent certainement être améliorées pour un contexte opérationnel.

4.5. Conclusion

Dans ce chapitre, nous avons procédé à la validation de notre méthodologie à l'aide de ShipMine. L'atelier a été utilisé pour extraire des exemples de motifs de mouvements et de situations de navires décrivant des comportements potentiellement à risques. Les tests ont été effectués sur des jeux de données réelles de déplacements et d'enquêtes d'accidents maritimes et ont permis de découvrir des connaissances sur des zones accidentogènes, des trajectoires de dérives, d'abordage, de pêche parallèle et des trajectoires type.

A ce stade de nos recherches, il est possible de répondre à notre problématique qui était de savoir si les motifs et les règles issus de la fouille de données peuvent décrire des comportements à risques. La fouille de données est donc appropriée pour la construction de connaissances sur les comportements à risques de navires.

Conclusion et perspectives

Conclusion

Nous nous sommes intéressés dans cette thèse à l'aide à la construction de connaissances sur les comportements à risques de navires. Les méthodes actuelles utilisées en amont de la modélisation des comportements à risques de navires sont souvent des méthodes d'acquisition de connaissances comme le brainstorming, les interviews d'experts et la revue de littérature. Les connaissances issues de ces méthodes ne sont pas nouvelles et dépendent beaucoup de l'expérience des experts. Dans ce travail de recherche, nous avons proposé une méthode originale d'aide à la construction de connaissances sur les comportements à risques de navires qui est basée sur la fouille de données. Cette fouille a permis de générer automatiquement des informations sur les comportements de navires à partir de l'exploration de données historiques de navigation et d'accidents.

La contribution de cette thèse peut être décrite en cinq points qui sont :

- **Proposition d'une nouvelle méthodologie de construction de connaissances sur les comportements à risques de navires,**

  La méthodologie proposée dans cette thèse est basée sur la fouille de données pour découvrir automatiquement des connaissances sous forme de motifs et de règles. Ces connaissances peuvent décrire des comportements à risques à partir des historiques de déplacements de navires et des événements maritimes à risques.

- **Identification de méthodes et algorithmes de fouille de données pouvant extraire des connaissances sur les comportements à risques de navires,**

  La problématique de cette thèse est de savoir si les connaissances issues de la fouille de données peuvent décrire des comportements à risques. Pour répondre à cette problématique, nous avons identifié des méthodes de fouille de données pouvant extraire ces comportements. Parmi ces méthodes, nous avons validé dans notre contexte, les méthodes de détection de facteurs de risques, de découverte de zones à risques, de trajectoires anormales et de navigations proches.

- **Constitution d'une base de données d'historiques de déplacements de navires, d'événements d'accidents maritimes britanniques et de météorologie,**

    - **Données spatiales statiques :** données MAIB (Marine Accident Investigation Branch), données MERRA (Modern-Era Retrospective analysis for Research and Applications),

    - **Données spatiales dynamiques :** Données AIS (Automatic Identification System).

- **Conception et développement d'un atelier de construction de connaissances sur les comportements à risques de navires basé sur la fouille de données,**

  Un atelier d'extraction de connaissances intégrant des méthodes de fouille de données et des historiques de données sur les navires a été proposé. Cet atelier peut aider des analystes, des experts maritimes et des scientifiques s'intéressant à l'analyse des comportements à risques de navires, à extraire des connaissances sur des comportements potentiellement à risques.

  L'atelier intègre des fonctionnalités d'extraction de zones à risques, de trajectoires aberrantes, de navigations proches et de routes de navigation maritimes.

- **Découverte de motifs de situations et de mouvements décrivant des comportements à risques,**

  Les méthodes de fouille de données intégrées à l'atelier, nous ont permis d'extraire quelques exemples de connaissances pouvant décrire des comportements à risques comme les zones accidentogènes, les trajectoires de dérives, d'abordage, de pêche parallèle et les routes potentiellement à risques.

- **Proposition d'une typologie de comportements potentiellement à risques de navires.**

  Cette typologie de comportements potentiellement à risques proposée à partir de l'analyse de la littérature, n'est pas exhaustive et peut être amenée à évoluer.

Perspectives

Nous proposons ci-après des perspectives pour améliorer l'atelier proposé, supporter la chaîne de traitement de l'information spatiale, améliorer la méthodologie et proposer une généralisation aux objets mobiles.

Amélioration de l'atelier

Deux perspectives à court terme sont envisagées pour améliorer l'atelier ShipMine. La première concerne l'extension de l'atelier en intégrant d'autres méthodes de fouille de données comme l'extraction de comportements périodiques (Cf. section 2.2.3.1.5 du chapitre 2) et le classement de trajectoires selon leurs comportements (Cf. section 2.2.3.1.3 du chapitre 2).

Les comportements périodiques semblent intéressants pour l'analyse des comportements à risques du fait qu'ils peuvent découvrir les comportements cycliques et leurs périodes, prédire les comportements habituels et détecter les comportements inhabituels qui s'écartent de ces comportements périodiques.

Le classement quant à lui, peut prédire la classe d'une trajectoire comme par exemple à risque à partir de caractéristiques décrivant ce type de classe. Pour ce faire, la méthode demande de disposer d'une base de données d'apprentissage où les trajectoires sont étiquetées comme trajectoires « à risques » ou « non à risques ». Cette méthode est difficile à mettre en œuvre du fait qu'il est difficile de trouver une telle base de données.

La seconde amélioration concerne la scalabilité (passage à l'échelle) de l'atelier par rapport à une avalanche de données. Dans ce travail de thèse, nous ne nous sommes pas penchés sur la question d'optimisation des algorithmes et de l'affichage dans ShipMine car les temps de réponses étaient acceptables pour les volumes de données explorées.

Couplage entre le data mining et le SOLAP

L'idée de couplage du *data mining* et de l'OLAP n'est pas nouvelle, J. Han (Han 1997) a proposé le concept de OLAP Mining à 1997. Ce concept est intéressant : il permet de sélectionner les portions de données pertinentes à partir du cube multidimensionnel[88] pour les explorer ; il offre la possibilité d'extraire des connaissances sur plusieurs niveaux d'abstraction en se basant sur des hiérarchies de concept et ; il présente les résultats de l'exploration via l'OLAP. Dans (Han & Kamber 2006), les auteurs parlent de l'utilisation du *data mining* avec l'OLAP pour une fouille interactive sur plusieurs niveaux d'abstraction en utilisant différents opérateurs de navigation : drilling, pivoting, filtering, dicing and slicing. Chacun de ces opérateurs multidimensionnels va permettre de générer un nouveau cube qu'il est possible d'explorer par *data mining*.

D'autres travaux de recherche relativement récents (Ben Messaoud et al., 2007) (Ramakrishnan & B.-C. Chen 2007) se sont intéressés aussi à ce concept de OLAP Mining. L'équipe de O. Boussaid (Ben Messaoud et al., 2007) par exemple, propose de faire de la fouille de données sur un environnement de données multidimensionnelles. Leur prototype OLEMAR permet d'extraire des règles d'association à partir d'un cube. Cela a permis d'éviter la répétition des règles car elles sont organisées en plusieurs niveaux d'abstraction.

L'analyse des phénomènes spatiaux a besoin d'une représentation cartographique pour étudier leur distribution spatiale. Par exemple, la partie 1 de la Figure 5-1, montre une représentation graphique et tabulaire d'un nombre d'accidents de navires par zone maritime. Ces représentations révèlent que le nombre d'accidents par zone est uniformément distribué mais la visualisation de la distribution des accidents sur une carte révèle des corrélations positives avec une trajectoire médiane de tankers (partie 2 de la Figure 5-1).

[88] C'est une méthode d'organisation et de représentation de données multidimensionnelles (Song and Miller, 2011). C'est une extension conceptuelle des tables de données relationnelles.

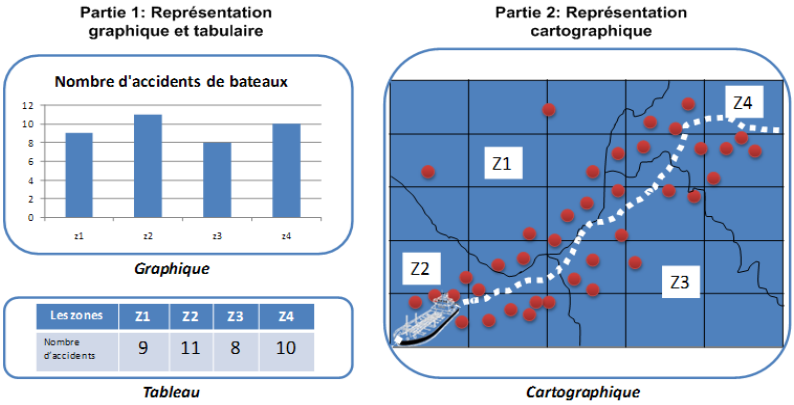


Figure 5-1 : Exemple de distribution des positions d'accidents maritimes par zone maritime.

Une perspective intéressante serait de savoir quel serait la pertinence du couplage du data mining et du SOLAP ? Que va apporter la dimension cartographique à la fouille de données ?

Le SOLAP comme outil de fouille de données visuelle

Le SOLAP a été défini par Yvan Bédard comme « *Une plateforme visuelle conçue spécialement pour supporter une analyse spatio-temporelle rapide et efficace à travers une approche multidimensionnelle qui comprend des niveaux d'agrégation cartographique, graphique et tabulaire* » (Bédard 1997).

Dans cette perspective, nous proposons une piste intéressante basée sur l'extraction de connaissances par fouille de données On-line (analyse multidimensionnelle). L'analyse multidimensionnelle peut être vue comme une approche semi-automatique par comparaison avec le *data mining* qui permet une exploration automatique des données. Dans cette fouille de données, ce sont les utilisateurs qui vont explorer visuellement les données spatio-temporelles pour en extraire du sens (relations, motifs).

Dans le domaine maritime, le SOLAP peut offrir une issue rapide et facile d'analyse des comportements de navires en mer. Les utilisateurs auront la possibilité de visualiser plusieurs indicateurs par croisement de plusieurs axes d'analyse (par exemple : type du navire, localisation relative, etc.) et sur plusieurs niveaux de granularités de détail. Le raisonnement des utilisateurs sur les différents croisements de données peut permettre la découverte visuelle de relations spatiales, non-spatiales et de motifs de mouvements.

Dans l'objectif de montrer qu'il est possible de découvrir des connaissances sur les situations à risques par analyse multidimensionnelle, nous présentons ci-après deux exemples, l'un sur l'extraction de relations et l'autre sur les zones à risques.

Pour l'extraction de relations entre des facteurs à risques, nous avons imaginé un cube multidimensionnel sur les accidents maritimes dont on présente ci-dessous sa structure (Figure 5-2). Dans ce schéma, le sujet d'analyse ou les faits sont les accidents maritimes qui sont étudiés par rapport aux mesures présentées dans la table « Mesures ». Cette table relie entre-elle les différentes dimensions (Types d'accidents, type de navire, temps, localisation, etc.) et comporte des mesures comme le nombre d'accidents, le nombre de morts avec leurs fonctions d'agrégations associées.

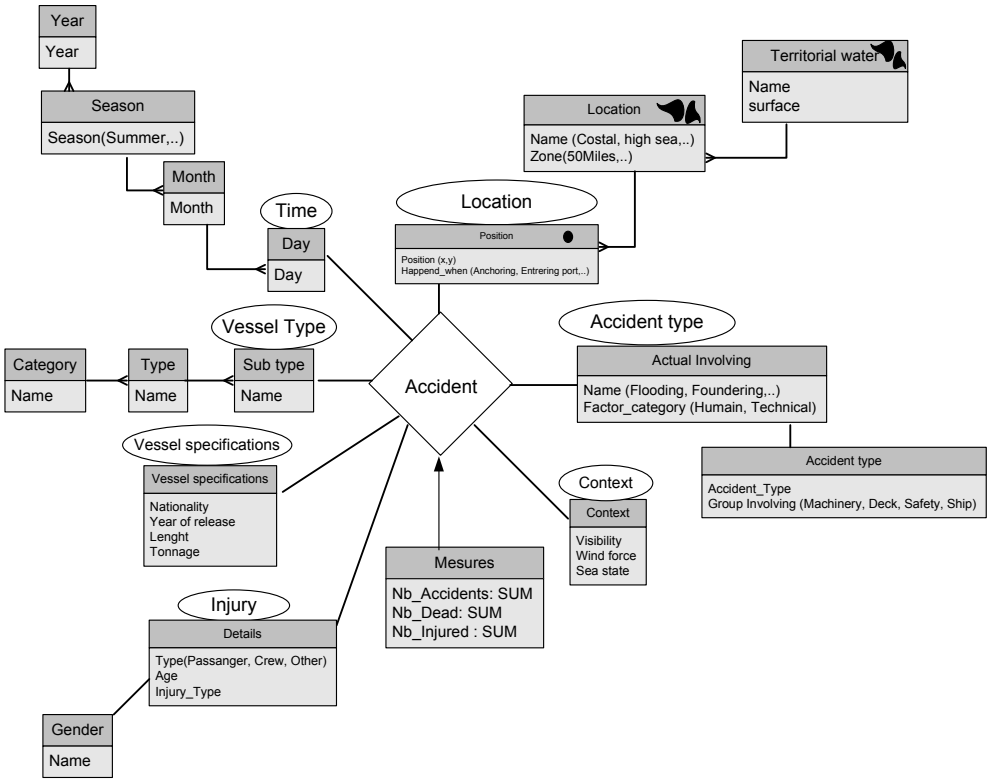


Figure 5-2 : Schéma en flocon du cube multidimensionnel des accidents maritimes.

L'utilisation d'un moteur *SOLAP* sur ce cube peut permettre à un utilisateur, par exemple d'afficher sur une cartographie le nombre d'accidents par eau territoriale et découvrir visuellement celles qui possèdent un nombre d'accidents très élevé. Cela peut permettre de découvrir des relations entre les accidents et les caractéristiques de ces eaux

territoriales. Ensuite, l'utilisateur peut se focaliser sur les eaux territoriales qui l'intéressent et changer de niveau d'abstraction en affichant cette fois, le nombre d'accidents par localisation. Donc il est possible d'extraire des situations à risques par analyse multidimensionnelle.

Concernant l'extraction de zones à risques, nous utilisons le résultat d'un stage effectué dans notre centre de recherche sur l'analyse multidimensionnelle de données de transport maritime. La réalisation d'un filtre sur les types de navires pour ne garder que ceux qui transportent des marchandises dangereuses va permettre la découverte de routes potentiellement à risques (Figure 5-3). Plus l'épaisseur de la trajectoire est grande, plus la quantité de marchandise dangereuse transportée est élevée. Pour avoir une meilleure représentation, il est possible de mettre des trajectoires types à la place des routes origine-destination.

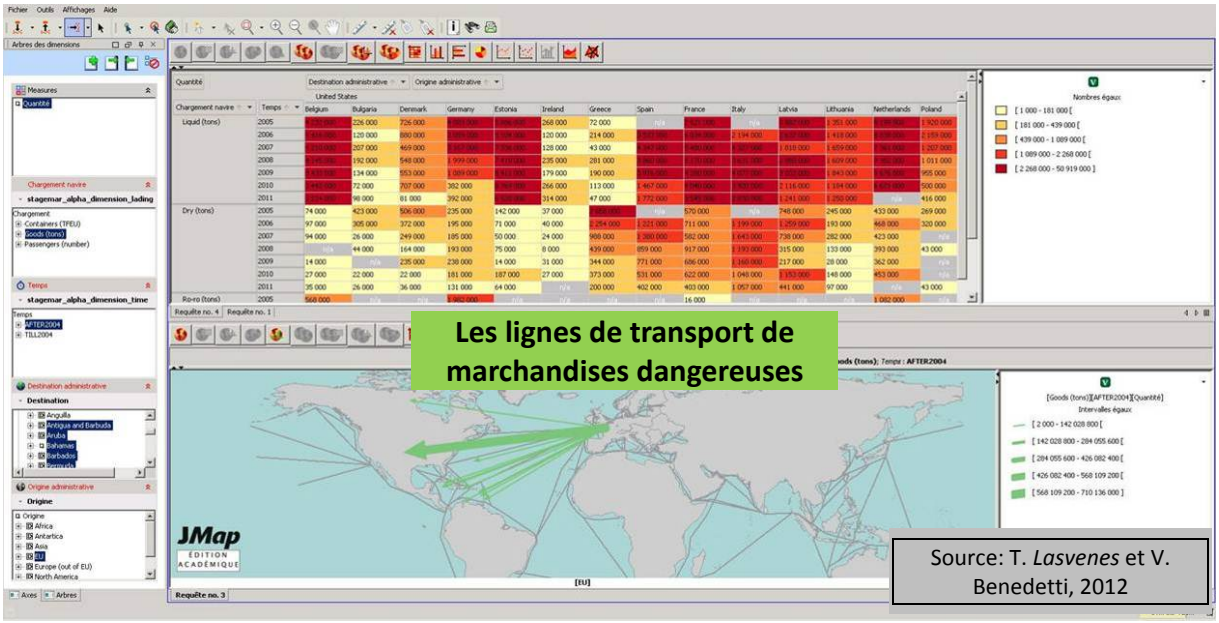


Figure 5-3 : Exemple de découverte de routes à risques par une analyse SOLAP

Les deux exemples d'extraction de connaissances présentés dans cette section, sont des situations à risques. Mais l'extraction de mouvements à risques est-elle possible à travers une analyse SOLAP ?

La réponse à cette question reste ouverte et les travaux sur les cubes multidimensionnels continus peuvent apporter des éléments de réponse. L'analyse multidimensionnelle conventionnelle ne supporte pas les données continues ce qui a soulevé un problème d'analyse des historiques de déplacement d'objets mobiles. En effet, les positions évoluent continuellement dans l'espace et le temps et les systèmes OLAP ne supportent pas les requêtes de sélection et d'agrégation les concernant. Pour améliorer cela, plusieurs travaux de recherches existent dans la littérature. Parmi ces travaux, nous pouvons citer ceux de l'équipe de Miller (Song & Miller 2011) qui proposent un cube de données continues pour les bases de données de flux de trafic et les travaux de l'équipe de Zeitouni (Wan et al., 2007)(Wan & Zeitouni 2005) qui ont étendu l'OLAP aux dimensions et aux faits continus pour les objets mobiles contraints par un réseau.

Utiliser l'OLAP et le SOLAP comme des outils de fouille de données présente quelques limites comme le temps nécessaire à la construction de connaissances. C'est dans ce cadre que nous proposons la chaîne géo-décisionnelle comme perspective de nos travaux.

La chaîne géo-décisionnelle

Nous proposons comme perspective, la définition d'une chaîne géo-décisionnelle (Figure 5-4 ci-dessous) qui va supporter la chaîne de traitement de l'information à des fins d'analyse. Par analogie à la chaîne décisionnelle, nous avons défini la chaîne géo-décisionnelle comme l'ensemble des outils formant la chaîne de traitement de l'information spatiale, de sa récupération à sa présentation aux décideurs (Idiri & Napoli 2013). La chaîne géo-décisionnelle est composée d'outils (spatial Extract Transform Load (ETL), Spatial Data Warehouse (SDW), SOLAP, spatial data mining (SDM), tableaux de bord (DB), etc.) méthodiquement assemblés pour assurer les 4 phases de l'informatique décisionnelle (collecte, stockage, distribution et exploitation), présentées parfois en 3 phases (alimentation, modélisation et analyse) (Fernandez 2010). Cette chaîne va

supporter toutes les fonctions décisionnelles de la collecte de données multi-sources à leur présentation aux utilisateurs finaux.

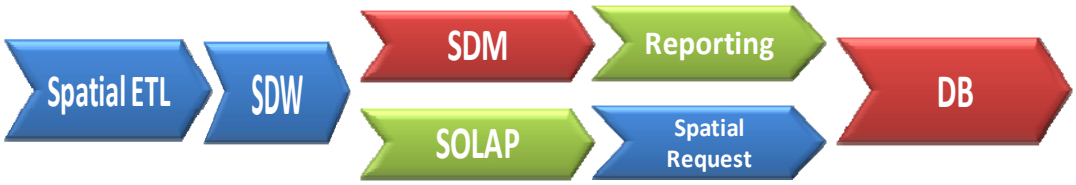


Figure 5-4 : La chaîne géo-décisionnelle.

La chaîne géodécisionnelle pousse à une réflexion d'ensemble où chaque composante doit remplir une tâche pour assurer une fonction globale qui est la production de l'information géo-décisionnelle. Les maillons qui composent la chaîne sont présentés ci-après :

Spatial Extract, Transform and Load (**Spatial ETL**) sont des outils permettant l'extraction de données ayant des dimensions géographiques à partir de sources hétérogènes (base de données, fichier plat, applications opérationnelles, etc.), leur transformation et leur chargement dans un entrepôt de données spatial (SDW). Ces outils ont la capacité de ré-exécuter à chaque mise à jour des sources de données ou à la demande les mêmes tâches qui ont permis la production d'informations décisionnelles.

Spatial Data Warehouse (SDW) est une base de données spatiale non volatile. Chaque donnée insérée dans cette base est estampillée pour pouvoir suivre son évolution au cours du temps. Cette base de données va constituer un référentiel unique pour tous les outils décisionnels qui vont venir interroger les données (*data mining*, SOLAP, etc.).

Spatial Data Mining (SDM) ensemble d'outils, méthodes et de techniques pour l'extraction non triviale de connaissances implicites et potentiellement utiles à partir des bases de données spatiales (Koperski & Han 1995).

Spatial OLAP (SOLAP) est un outil permettant une analyse multidimensionnelle des données spatiales sur plusieurs niveaux d'agrégation.

Reporting est un regroupement d'outils d'élaboration automatique de rapports d'aide à la décision ou des états de sorties à partir des données sauvegardées dans l'entrepôt de données ou de résultats obtenus après analyse.

Spatial request est une interface de requêtage intuitif permettant d'interroger l'entrepôt de données (SDW).

Dashboard permet le pilotage d'une activité par le suivi d'indicateurs clés.

Identification temps-réel des comportements à risques

Nous avons comme perspective la classification et l'exploitation des connaissances découvertes par *data mining*.

La classification des connaissances peut permettre la constitution de comportements à risques en croisant les connaissances des situations et mouvements à risques par une approche matricielle. Cela va permettre d'obtenir une catégorisation qui va spécifier les comportements à risques.

Pour automatiser l'identification à temps des risques, une approche de raisonnement automatique à base de règles est pertinente (Idiri & Napoli 2012b). Sa modularité sous forme de règles de connaissances rend simple sa maintenabilité. De plus, il est facile et rapide d'ajouter ou supprimer des connaissances.

Le raisonnement automatique à base de règles peut permettre de tester et valider les connaissances précédemment découvertes sur des flux de données réelles de déplacements de navires.

Généralisation de la méthodologie aux objets mobiles

La généralisation de notre méthodologie au domaine des transports ou aux objets mobiles peut être une perspective intéressante. En effet, il y a aujourd'hui d'énormes quantités de données décrivant des déplacements de mobiles (avions, navires, personnes, etc.) qui sont stockées. Ces données sont la plupart du temps exploitées pour la surveillance temps-réel. L'une des problématiques de la surveillance aujourd'hui est le développement de méthodes efficaces pour découvrir automatiquement les risques à partir de ces données de déplacements et d'autres données complémentaires comme le contexte, l'environnement et le renseignement.

Bibliographie

Agrawal, R., Imielinski, T. & Swami, A., 1993. Mining Association Rules between Sets of Items in Large Databases. In *International Conference on Management of Data (ACM SIGMOD)*. Washington, D.C., USA, pp. 207–216.

Agrawal, R. & Srikant, R., 1994. Fast Algorithms for Mining Association Rules. In *the 20th International Conference on Very Large Data Bases*. Santiago, Chile, pp. 487–499.

Amrozowicz, M.D., Brown, A. & Golay, M., 1997. A Probabilistic Analysis Of Tanker Groundings. In *7th International Offshore and Polar Engineering Conference*. Honolulu, Hawaii, pp. 1–19.

Anderberg, M.R., 1973. *Cluster Analysis for Applications (Probability & Mathematical Statistics Monograph)* Academic Press Inc, ed., United Kingdom.

Andrienko, G., Andrienko, N., Rinzivillo, S., Nanni, M., Pedreschi, D. & Giannotti, F., 2009. Interactive visual clustering of large collections of trajectories. In *IEEE Symposium on Visual Analytics Science and Technology (VAST 2009)*. Atlantic City, New Jersey, USA, pp. 3–10.

Andrienko, G., Andrienko, N., Schumann, H. & Tominski, C., 2014. Visualization of Trajectory Attributes in Space–Time Cube and Trajectory Wall. In M. Buchroithner, N. Prechtel, & D. Burghardt, eds. *Lecture Notes in Geoinformation and Cartography*. Springer Berlin Heidelberg, pp. 157–163.

Aufaure, M., Yeh, L. & Zeitouni, K., 2000. *Le temps, L'espace et l'évolutif en sciences du traitement de l'information - Tome 2* Cepadues.,

Autran, O., 2011. Maritime Surveillance cost - Benefit analysis Key Findings. In *SeaTIMed -SPACEMAR*. Toulon, France.

Awasthi, A., Lechevallier, Y., Parent, M. & Proth, J.M., 2005. Rule based prediction of fastest paths on urban networks. In *Intelligent Transportation Systems. IEEE Proceedings*. pp. 978–983.

Bédard, Y., 1997. Spatial OLAP. In *Forum annuel sur la R-D, Géomatique VI: Un monde accessible*. pp. 13–14.

Bellayer Roille, A., 2011. Les enjeux politiques autour des frontières maritimes. *CERISCOPE Frontières*. Available at: http://ceriscope.sciences-po.fr/node/15.

Ben Messaoud, R., Rabaséda, S.L., Missaoui, R. & Boussaid, O., 2007. OLEMAR: an On-Line Environment for Mining Association Rules in Multidimensional Data. In *Data Advances in Data Warehousing and Mining*. Idea Group Inc, pp. 14–47.

Benkert, M. et al., 2008. Reporting flock patterns. *Computational Geometry*, 41(3), pp.111–125.

Berche, P., 2007. *Une histoire des microbes* John Libbe., United Kingdom.

Bodewig, K., Kouts, T. & Konstantinos, V., 2009. *La surveillance maritime européenne*, Available at: http://www.assembly-weu.org/fr/documents/sessions_ordinaires/rpt/2009/2051.pdf.

Boisson, P., 1998. *Politiques et droit de la sécurité maritime.* Bureau Veritas., Paris.

Breiman, L., Friedman, J., Olshen, R. & Stone, C., 1984. *Classification and Regression Trees.* Chapman & Hall/CRC, ed.,

Breunig, M., Kriegel, H.P., T. Ng, R. & Sander, J., 2000. LOF: Identifying Density-Based Local Outliers. In *Proceedings of the ACM SIGMOD. International Conference on Management of Data.* New York, NY, USA: ACM, pp. 93–104.

Britz, C., 2011. Spationav : la bande côtière sous haute surveillance. *Le Marin*.

Buard, E. & Brasebin, M., 2011. Visual exploration of large animal trajectories. In *25th International Cartographic Conference (ICC 2011)*. Paris, France.

Buard, E. & Christophe, S., 2012. Towards a data model of human and wildlife trajectories: describing individuals' movements with the help of Time-Geography concepts. In *Workshop on Time Geography*. Colombus, USA.

Cao, H., Wolfson, O. & Trajcevski, G., 2006. Spatio-temporal Data Reduction with Deterministic Error Bounds. *The VLDB Journal*, 15(3), pp.211–228.

Cao, H., Mamoulis, N. & Cheung, D., 2007. Discovery of Periodic Patterns in Spatiotemporal Sequences. *IEEE Transactions on Knowledge and Data Engineering*, 19(4), pp.453–467.

Chaze, X., Bouejla, A., Napoli, A., Guarnieri, F., Eude, T. & Alhadef, B., 2012. The Contribution of Bayesian Networks to Manage Risks of Maritime Piracy against Oil Offshore Fields. In *International Workshops Information Technologies for the Maritime Sector (ITEMS 2012), Database Systems for Advanced Applications, Lecture Notes in Computer Science Volume 7240*. Busan, South-Korea: Springer Berlin Heidelberg, pp. 81–91.

Chelghoum, N., Zeitouni, K., Laugier, T., Fiandrino, A. & Loubersac, L., 2006. Fouille de données spatiales Approche basée sur la programmation logique inductive. In *6èmes Journées d'Extraction et de Gestion des Connaissances, Edition CEPADUES*. Lille, France.

Chelghoum, N. & Zeitouni, K., 2004. Data mining spatial un probleme de data mining multi-tables. *RIST*, 14(2), pp.129–145.

Chen, J., Lai, C., Meng, X., Xu, J., & Hu, H., 2007. Clustering moving objects in spatial networks. In *Proceedings of the 12th international conference on Database systems for advanced applications*. Berlin, Heidelberg: Springer-Verlag, pp. 611–623.

CNUCED, 2009. *Etude sur les transports maritimes*, Geneva. Available at: http://www.unctad.org/fr/docs/rmt2009_fr.pdf.

Committee, 2003. *Livre IT Roadap to a geospatial future* The Nation., Washington, D.C.

Cougnaud, B., 2007. *L'univers des risques en finance* Les Presses de Sciences PO, ed.,

Darpa, 2005. Predictive Analysis For Naval Deployment Activities (PANDA). Available at: https://www.fbo.gov/index?s=opportunity&mode=form&tab=core&id=7c1de14f8cdfd3d14f112f99a2bbd782&_cview=0.

Deiss, H., 2011. Sûreté maritime : la piraterie a progressé de 86% en cinq ans. Available at: http://www.wk-transport-logistique.fr/actualites/detail/35793/surete-maritime-la-piraterie-a-progresse-de-86-en-cinq-ans.html [Accessed August 10, 2011].

Dhafer, L., Boullé, M. & Laurent, D., 2012. Prétraitement Supervisé des Variables Numériques pour la Fouille de Données Multi-Tables. In *Extraction et Gestion des Connaissances*. Bordeaux, France.

El Moussawi, A., 2013. *Développement d'un prototype d'analyse de comportements de navires basé sur les algorithmes de fouille de données.* Centre de recherche sur les Risques et les Crises (CRC) - MINES ParisTech.

Ester, M., Kriegel, H.P., Sander, J. & Xu, X., 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In E. Simoudis, J. Han, & U. M. Fayyad, eds. *International Conference on Knowledge Discovery and Data Mining (KDD-96)*. Portland, Oregon, USA: AAAI Press, pp. 226–231.

Ester, M., Kriegel, H.P. & Sander, J., 1997. Spatial data mining: A database approach. *Advances in Spatial Databases*, 1262, pp.47–66.

Estevez, I., Dubois, S., Gartiser, N., Renaud, J. & Caillaud, E., 2006. Le raisonnement à partir de cas est-il utilisable pour l'aide à la conception inventive. In *14e Atelier de Raisonnement à Partir de Cas*. Besançon, France, pp. 30–31.

Etienne, L., 2011. *Motifs spatio-temporels de trajectoires d'objets mobiles, de l'extraction à la détection de comportements inhabituels. Application au trafic maritime.* UNIVERSITÉ DE BRETAGNE OCCIDENTALE.

Etienne, L., Devogele, T. & Bouju, A., 2008. Outils d'aide aux décideurs concernant le suivi de navires : suivi de trajectoires relatives entre navires et détection de trajectoires inhabituelles. In *7ème journées scientifiques et techniques du CETMEF*. Paris, France.

Etienne, L., Devogele, T. & Bouju, A., 2009. Analyse de similarité de trajectoires d'objets mobiles suivant le même itinéraire: Application aux trajectoires de navires. *Ingénierie des Systèmes d'Information*, 14(5/2009), pp.85–106.

Etienne, L., Devogele, T. & Bouju, A., 2010. Spatio-temporal trajectory analysis of mobile objects following. In *the International Symposium on Spatial Data Handling (SDH)*. Hong Kong, pp. 86–91.

Etienne, L., Ray, C. & McARDLE, G., 2011. Spatio-Temporal Visualisation of Outliers. In *First international workshop on Maritime Anomaly Detection (MAD 2011)*. Tilburg University, The Netherlands, pp. 19–20.

FAO, 2010. *Situation mondiale des pêches et de l' aquaculture*, Available at: http://www.fao.org/docrep/013/i1820f/i1820f01.pdf.

Fernandez, A., 2010. *Les nouveaux tableaux de bord des managers* 5th ed., Editions d'Organisation.

Fournier, S., 2005. *Intégration de la dimension spatiale au sein d'un modèle multi-agents à base de rôles pour la simulation : Application à la navigation maritime*. l'Université de Rennes 1, France.

Frawley, W.J., Piatetsky-shapiro, G. & Matheus, C.J., 1992. in Databases: An Overview. *AI Magazine*, 13(3), pp.57–70. Available at: http://www.aaai.org/ojs/index.php/aimagazine/article/view/1011.

Gaffney, S.J., Robertson, A.W., Smyth, P., Camargo, S.J., & Ghil, M., 2006. *Probabilistic Clustering of Extratropical Cyclones Using Regression Mixture Models*, University of California, Irvine, USA.

Gardarin, G., 2011. *Bases de données. Objet & relationnel* Eyrolles, ed., Available at: http://www.decitre.fr/livres/bases-de-donnees-9782212092837.html.

Gardarin, G., Pucheral, P. & Wu, F., 1998. Bitmap Based Algorithms For Mining Association Rules. In *proceeding of 14ème Journées Bases de Données Avancées*. Hammamet, Tunisie, pp. 1–19.

Giannotti, F., Nanni, M., Pinelli, F. & Pedreschi, D., 2007. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD' 07)*. New York, NY, USA, pp. 330–339.

Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S. & Trasarti, R., 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, 20(5), pp.695–719.

Giraud, M., Alhadef, B., Chaze, X., Napoli, A., Naudin, A.C, Bottala-gambetta, M., Grimaldi, G., Chaumartin, D., Morel, M., Imbert, C., Wasselin, J.P., Bonacci, D.

& Michel, P., 2013. «SARGOS» Système d'Alerte et Réponse Graduée Off Shore. In *Conférence WISG 2013 - Workshop Interdisciplinaire sur la Sécurité Globale*. Troyes, France.

Gouin, D., Lavigne, V. & Davenport, M., 2011. Towards Collaborative Visual Analytics of a Vessel of Interest. In *The Third International UKVAC Workshop on Visual Analytics VAW2011*. London.

Graham, R. & Yao, F., 1983. Finding the Convex Hull of a Simple Polygon. *Journal of Algorithms*, 4, pp.324–331.

Grünwald, P., Myung, J. I., & Pitt, M., 2005. Advances in Minimum Description Length: Theory and Applications, p. 372. MIT Press.

Gudmundsson, J. & Van Kreveld, M., 2006. Computing longest duration flocks in trajectory data. In *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*. New York, NY, USA, pp. 35–42.

Han, J., 1997. OLAP Mining: An Integration of OLAP with Data Mining. In *Proceedings of the 7th IFIP 2.6 Working Conference on Database Semantics (DS-7)*. pp. 3–20.

Han, J., Krzysztof, K., Nebojsa, S., 1997. GeoMiner: a system prototype for spatial data mining. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, pp. 553–556.

Han, J., Pei, J., Yin, Y., 2000. Mining frequent patterns without candidate generation. In *SIGMOD '00 Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. New York, NY, USA, pp. 121–5.

Han, J. & Kamber, M., 2006. *Data mining concepts and technique* Second Edi. T. M. K. S. I. D. M. A. Systems, ed.,

Heredia, J.M.S., 2009. L'approche nationale en matière des zones maritimes en Méditerranée. *Anuario da Facultade de Dereito da Universidade da Coruña*, 13, pp.753–772.

Idiri, B. & Napoli, A., 2012a. Découverte de règles d'association pour l'aide à la prévision des accidents maritimes. In *12ème Conférence Internationale Francophone sur l'Extraction et la Gestion de Connaissance (EGC 2012), Revue des Nouvelles Technologies de l'Information RNTI-E-23 Éditions Hermann*. Bordeaux, France, pp. 243–248.

Idiri, B. & Napoli, A., 2012b. The automatic identification system of maritime accident risk using rule-based reasoning. In *IEEE Internationale Conference on System of Systems Engineering (SOSE 2012)*. Genoa, Italie.

Idiri, B. & Napoli, A., 2013. The geographical decision-making chain: formalization and application to maritime risk analysis. In *International Workshop on*

Information Fusion and Geographic Information Systems: Environmental and Urban Challenges (IF&GIS' 2013). St. Petersburg, Russie.

Jami, S., Jen, T.Y., Laurent, D., Loizou, G. & Sy, O., 2005. Extraction de règles d'association pour la prédiction de valeurs manquantes. *Revue africaine de la recherche en informatique et mathématiques appliquées (AREMA)*, 3(numéro spécial CARI'04), pp.103–124.

Jeung, H., Yiu, M.L., Zhou, X., Jensen, C.S. & Shen, H.T., 2008a. Discovery of convoys in trajectory databases. *Proc. VLDB Endow.*, 1(1), pp.1068–1080.

Jeung, H., Shen, H.T. & Zhou, X., 2008b. Convoy Queries in Spatio-Temporal Databases. *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pp.1457–1459.

Jones, H., 2007. *raisonnement à base de règles implicatives floues -Inférence et Sémantique-*. Université de Touleuse III, Paul Sabatier.

Kai-Lin, P., Guan-Hong, C. & Wei-Guang, T., 2013. Discovering unusual behavior patterns from motion data. *IEEE International Conference on Consumer Electronics (ICCE)*, pp.242 – 243.

Kalnis, P., Mamoulis, N. & Bakiras, S., 2005. On Discovering Moving Clusters in Spatio-temporal Data. In *Proceedings of the 9th International Conference on Advances in Spatial and Temporal Databases*. Berlin, Heidelberg: Springer-Verlag, pp. 364–381.

Kharrat, A., Zeitouni, K., Popa, I.S. & Faiz, S., 2008a. Clustering Algorithm for Network Constraint Trajectories. In *13th International Symposium on Spatial Data Handling*. pp. 1–17.

Kharrat, A., Zeitouni, K. & Faiz, S., 2008b. Clustering de trajectoires contraintes par un réseau. In *Atelier Fouille de données complexes, 8èmes Journées d'Extraction et de Gestion des Connaissances*. Sophia-Antipolis, France.

Kharrat, A., Popa, I.S., Zeitouni, K. & Faiz, S., 2009. Caractérisation de la densité de trafic et de son évolution à partir de trajectoires d'objets mobiles. In *Proceedings of the 5th French-Speaking Conference on Mobility and Ubiquity Computing*. New York, NY, USA: ACM, pp. 33–40.

Kharrat, A., Zeitouni, K. & Faiz, S., 2013. Découverte de patrons de mobilité dans un réseau. In *Atelier de FOuille de données Spatio-Temporelles et Application (FOSTA) - EGC2013*. Toulous, France, pp. 13–24.

Koperski, K. & Han, J., 1995. Discovery of Spatial Association Rules in Geographic Information Databases. In M. J. Egenhofer & J. R. Herring, eds. *Advances in Spatial Databases: 4th International Symposium SSD'95*. Portland, ME, USA: Springer-Verlag, pp. 47–66.

Ladiray, D., 1997. L'analyse exploratoire des données. *Courrier des Statistiques*, pp.3–6. Available at: http://www.insee.fr/fr/ffc/docs_ffc/cs90a.pdf.

Laere, J. Van & Nilsson, M., 2009. Evaluation of a workshop to capture knowledge from subject matter experts in maritime surveillance. In *12th International Information Fusion Conference*. Seattle, USA, pp. 171–178.

Lai, C., Wang, L., Chen, J., Meng, X. & Zeitouni, K., 2007. Effective density queries for moving objects in road networks. In *APWeb/WAIM'07 Proceedings of the joint 9th Asia-Pacific web and 8th international conference on web-age information management conference on Advances in data and web management*. Berlin, Heidelberg: Springer-Verlag, pp. 200–211.

Lavigne, V. & Gouin, D., 2011. *Visual Analytics for Defence and Security*, Valcartier.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. & Alstyne, M.V., 2009. Computational Social Science. *Science*, 323(5915), pp.721–723.

Le Bras, Y., Lallich, S. & Lenca, P., 2011. Un cadre formel pour l'étude des mesures d'intérêt des réegles d'association. In *Journée d'animation du GRD I3 sur la fouille de données*. Lyon, France.

Le Pors, T., Devogele, T. & Chauvin, C., 2009. Multi agent system integrating Naturalistic Decision Roles: application to maritime traffic. In *IADIS International Conference Intelligent Systems and Agents*. Portugal.

LeBlanc, L.A. & Rucks, C.T., 1996. A multiple discriminant analysis accidents. *Accident Analysis & Prevention*, 28(04), pp.501–510.

Lee, J.G., Han, J. & Whang, K.Y., 2007. Trajectory clustering: a partition-and-group framework. In *Proceedings of the ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, pp. 593–604.

Lee, J.G., Han, J. & Li, X., 2008a. Trajectory Outlier Detection: A Partition-and-Detect Framework. In *Data Engineering, IEEE International Conference on Data Engineering (ICDE'2008)*. Cancun, Mexique, pp. 140–149.

Lee, J.G., Han, J., Li, X. & Gonzalez, H., 2008b. TraClass: trajectory classification using hierarchical region-based and trajectory-based clustering. *Proc. VLDB Endow.*, 1(1), pp.1081–1094.

Li, Z., Ding, B., Han, J. & Kays, R., 2010a. Swarm: mining relaxed temporal moving object clusters. *Proc. VLDB Endow.*, 3(1-2), pp.723–734.

Li, Z., Ding, B., Han, J., Kays, R. & Nye, P., 2010b. Mining periodic behaviors for moving objects. In ACM, ed. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'2010)*. New York, NY, USA, pp. 1099–1108.

Li, Z., Ji, M., Lee, J.G., Tang, L.A. & Han, J., 2010c. MoveMine: Mining Moving Object Databases. In *International Conference On Management of Data (SIGMOD)*. Indianapolis.

Li, Z., Han, J., Ji, M., Tang, L.A.,Yu, Y., Ding, B., Lee, J.G. & Kays, R., 2011. MoveMine: Mining moving object data for discovery of animal movement patterns. *ACM Trans. Intell. Syst. Technol.*, 2(4), pp.37:1–37:32.

Lieber, J., 2001. Des règles, des cas, des généralités, des spécificités, des applications, des adaptations, des chaînes, des combinaisons et des tartes. In A. Mille & B. Fuchs, eds. *Atelier raisonnement à partir de cas RàPC 2001*. Grenoble, France, pp. 49–58.

Lu, C.-T., Boedihardjo, A.P. & Zheng, J., 2006. AITVS: Advanced Interactive Traffic Visualization System. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*. p. 167.

Martin, O., 2012. Induction-déduction. *Sociologie [En ligne], Les 100 mots de la sociologie*, pp.13–14. Available at: http://sociologie.revues.org/1594.

Marven, C., Canessa, R. & Keller, P., 2007. Exploratory Spatial Data Analysis to Support Maritime Search and Rescue Planning. In J. Li, S. Zlatanova, & A. G. Fabbri, eds. *Geomatics Solutions for Disaster Management*. New York: Springer Berlin Heidelberg, pp. 271–288.

Masseglia, F., Teisseire, M. & Poncelet, P., 2004. Extraction de motifs séquentiels Problèmes et méthodes. *Ingénierie des Systèmes d'Information*, pp.183–210.

Meratnia, N. & De By, R.A., 2004. Spatiotemporal Compression Techniques for Moving Point Objects. In E. Bertino et al., eds. *Advances in Database Technology - 9th International Conference on Extending Database Technology*. Heraklion, Crete, Greece: Springer Berlin Heidelberg, pp. 765–782.

Miller, H.J. & Han J., 2009. *Geographic Data Mining and Knowledge Discovery, Second Edition* K. Vipin, ed., Minneapolis, Minnesota, USA: Press, CRC.

Morel, M., 2009. SisMaris : Système d'Information et de Surveillance MARitime pour l'Identification des comportements Suspects de navire. In *première Conférence Méditerranéenne Côtière et Maritime CM 2*. Hammamet, Tunisie, pp. 261–264.

Morel, M., Napoli, A., Littaye, A., Georgé, J.P. & Jangal, F., 2008. Surveillance et contrôle des activités des navires en mer. In *Workshop Interdisciplinaire sur la Sécurité Globale (WISG08)*. Troyes, France.

Morel, M., Napoli, A., Georgé, J.P., Jangal, F., Giraud, M. & Bottala-gambetta, M., 2010. Surveillance et controle des activites des navires en mer ScanMaris. In *Workshop Interdisciplinaire sur la Sécurité Globale - WISG 2010*. Troyes, France, pp. 1–10.

Morel, M. & Broussolle, J., 2011. I2C, Interoperable sensors & information sources for Common detection of abnormal vessel behaviours and collaborative suspect events analysis. In *MAST 2011 Conference Session*. Marseille, France: MAST Events Ltd.

Morel, M., Flori, V., Poirel, O., Napoli, A., Salom, P. & Proutiere Maulion, G., 2011. Trairtement et Authentification des MenAces et RISques en mer. In Worksh*op Interdisciplinaire sur la Sécurité Globale (WISG11)*. Troyes, France.

Nathan, R, Getz, W.M., Revilla, E., Holyoak, M., Kadmon, R., Saltz, D. & Smouse, P.E., 2008. A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 105(49), pp.19052–19059.

Nilsson, M., Laere, J.V., Ziemke, T. & Edlund, J., 2008. Extracting rules from expert operators to support situation awareness in maritime surveillance. In *11th International Information Fusion Conference*. Cologne, Germany, pp. 1–8.

Nishizaki, C., Itoh, H., Yoshimura, K., Hikida, K. & Mitomo, N., 2011. Development of a method for marine accident analysis with bridge simulator. In *System of Systems Engineering (SoSE), 6th International Conference on on System of Systems Engineering*. Albuquerque, New Mexico, USA: IEEE, pp. 31–36.

Noyon, V., 2007. *Modélisation de vue relative et spatio-temporelle de trajectoires*. l'Ecole Nationale Supérieur d'Arts et Métiers.

OMI, 2009. OMI Ce qu'elle est. , pp.24–49. Available at: http://www.imo.org/About/Documents/IMO What-it-is web 2009.pdf [Accessed October 25, 2013].

Ong, R., Pinelli, F., Trasarti, R., Nanni, M., Renso, C., Rinzivillo, S. & Giannotti, F., 2011. Traffic Jams Detection Using Flock Mining. In *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III*. Berlin, Heidelberg: Springer-Verlag, pp. 650–653.

Parsaye, K., 1995. The Sandwich Paradigm. *Database Programming & Design*, pp.50–55.

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U. & Hsu, M.C., 2001. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In *Proceedings of 17th International Conference on Data Engineering (ICDE'01)*.

Prati, Y., 2010. Piloter en sûreté et à bon port la sécurité nationale. *Les cahiers thématiques de la Revue Défence Nationale*, p.315. Available at: http://www.defnat.com/site_fr/pdf/TRIBUNE-220910-PRATI.pdf.

Quinlan, J.R., 1979. Discovering rules by induction from large collections of examples. In D. Michie, ed. *Expert systems in the micro electronic age*. Edinburgh: Edinburgh University Press, pp. 168–201.

Quinlan, J.R., 1982. Semi-autonomous acquisition of pattern-based knowledge. In J. . Hayes, D. Michie, & Y.-H. Pao, eds. *Machine intelligence 10*. Chichester: Ellis Horwood, pp. 159–172.

Quinlan, J.R., 1986. Induction of Decision Trees. *Machine Learning*, 1(1), pp.81–106.

Ramakrishnan, R. & Chen, B.-C., 2007. Exploratory mining in cube space. *Data Mining and Knowledge Discovery*, 15(1), pp.29–54.

Riveiro, M., Falkman, Göran & Ziemke, T., 2008. Improving maritime anomaly detection and situation awareness through interactive visualization. *11th International Conference on Information Fusion*, pp.1–8.

Riveiro, M. & Falkman, Göran, 2009. Interactive Visualization of Normal Behavioral Models and Expert Rules for Maritime Anomaly Detection. In *International Conference on Computer Graphics, Imaging and Visualization*. Los Alamitos, CA, USA: IEEE Computer Society, pp. 459–466.

Riveiro, M. & Falkman, Göran, 2011. The role of visualization and interaction in maritime anomaly detection P. Chung Wong et al., eds. *Proceedings of SPIE, the International Society for Optical Engineering*, 7868.

Roddick, J.F., Hornsby, K. & Spiliopoulou, M., 2001. A bibliography of temporal, spatial and spatio-temporal data mining research. *ACM SIGKDD Explorations Newsletter*, 2007/2001, pp.147–163.

Roddick, J.F. & Lees, B.G., 2009. Paradigms for spatial and spatio-temporal data mining. In H. Miller & J. Han, eds. *Geographic Data Mining and Knowledge*. New York, USA: CRC Press, pp. 27–44.

Roy, J., 2008. Anomaly detection in the maritime domain. In *Proceeding of Optics and Photonics in Global Homeland Security IV, 69450W*. Spie.

Roy, J., 2010. Rule-based expert system for maritime anomaly detection. In *SPIE, the International Society for Optical Engineering*. Orlando, Florida, USA, p. 12.

Silverstein, C., Brin, S. & Motwani, R., 1998. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2(1), pp.39–68.

Song, Y. & Miller, H., 2011. Exploring traffic flow databases using space-time plots and data cubes. *Transportation*, 39, pp.215–234.

Srikant, R. & Agrawal, R., 1996. Mining Sequential Patterns□: Generalizations and Performance Improvements. In *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology (EDBT'96)*. London, UK: Springer-Verlag, pp. 3–17.

Studer, R., Benjamins, V.R. & Fensel, D., 1998. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1-2), pp.161–197.

Sumpter, N. & J. Bulpitt, A., 2000. Learning Spatio-Temporal Patterns for Predicting Object Behaviour. *Image Vision and Computing*, 18(9), pp.697–704.

TéSA, 2011. *Livrable de projet SARGOS ANR-09SECU-09. Classification de cibles marines a partir de l'echo radar*,

Torun, A. & Düzgün, S., 2006. Using spatial data mining techniques to reveal vulnerability of people and places due to oil transportation and accidents: a case study of istanbul strait. In *ISPRS Technical Commission II Symposium*. Vienna, pp. 12 – 14.

Tufféry, S., 2010. *Data mining et statistique décisionnelle* 3rd ed. TECHNIP, ed.,

Tukey, J.W., 1980. We Need Both Exploratory and Confirmatory. *The American Statistician*, 34(1), pp.23–25.

Tung, A., Hou, J. & Han, J., 2001. Spatial clustering in the presence of obstacles. In *Proceedings of 17th International Conference on Data Engineering*. pp. 359–367.

Vaillant, B., Meyer, P., Prudhomme, E., Lallich, S., Lenca, P. & Bigaret, S., 2005. Mesurer l'intérêt des règles d'association. In *Atelier Qualité des Données et des Connaissances (EGC 2005)*. Paris, France, pp. 69–78.

Vandecasteele, A., 2012. *Modélisation ontologique des connaissances expertes pour l'analyse de comportements à risque. Application à la surveillance maritime*. Ecole nationale supérieure des MINES de Paris.

Vasyechko, O.A., Benlagha, N. & Grun-Rehomme, M., 2005. Comparaison de méthodes de détection des valeurs extrêmes: Application en statistique d'entreprise. Available at: http://econpapers.repec.org/RePEc:erm:papers:0603.

Vatin, G. & Napoli, A., 2013a. Guiding the Controller in Geovisual Analytics to Improve Maritime Surveillance. In *The Fifth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2013)*. Nice, France, pp. 26–31.

Vatin, G. & Napoli, A., 2013b. Toward User-Centred Geovisual Analytics in Maritime Surveillance. In *GeoViz*. Hamburg, Allemagne.

Vieira, M.R., Bakalov, P. & Tsotras, V.J., 2009. On-line discovery of flock patterns in spatio-temporal data. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. New York, NY, USA: ACM, pp. 286–295.

Wan, T. & Zeitouni, K., 2005. Modélisation d'objets mobiles dans un entrepôt de données. In *EGC'05*. pp. 343–348.

Wan, T., Zeitouni, K. & Meng, X., 2007. An OLAP system for network-constrained moving objects. In *Proceedings of the 2007 ACM symposium on Applied computing*. New York, NY, USA: ACM, pp. 13–18.

Willems, N., 2011. *Visualization of Vessel Traffic*. Eindhoven University of Technology.

Willems, N., Van de Wetering, H. & Van de Wijk, J.J., 2009. Visualization of vessel movements. *Computer Graphics Forum*, 28(3), pp.959–966.

Willems, N., Van de Wetering, H. & Van Wijk, J.J., 2011. Evaluation of the Visibility of Vessel Movement Features in Trajectory Visualizations. *Computer Graphics Forum*, 30(3), pp.801–810.

Zaki, M.J., 2001. SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Mach. Learn.*, 42(1-2), pp.31–60.

Zhang, Z. & Feng, X., 2009. New Methods for Deviation-Based Outlier Detection in Large Database. In *Sixth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'09)*. pp. 495–499.

Méthodologie d'extraction de connaissances spatio-temporelles par fouille de données pour l'analyse de comportements à risques - Application à la surveillance maritime

RESUME :

Les progrès technologiques en systèmes de localisation (AIS, radar, GPS, RFID, etc.), de télétransmission (VHF, satellite, GSM, etc.), en systèmes embarqués et leur faible coût de production ont permis leur déploiement à une large échelle. Enormément de données sur les déplacements d'objets sont produites par le biais de ces technologies et utilisées dans diverses applications de surveillance temps-réel comme la surveillance du trafic maritime. L'analyse a posteriori des données de déplacement de navires et d'événements à risques peut présenter des perspectives intéressantes pour la compréhension et l'aide à la modélisation des comportements à risques. Dans ce travail de thèse une méthodologie basée sur la fouille de données spatio-temporelle est proposée pour l'extraction de connaissances sur les comportements potentiellement à risques de navires. Un atelier d'aide à l'analyse de comportements de navires fondé sur cette méthodologie est aussi proposé.

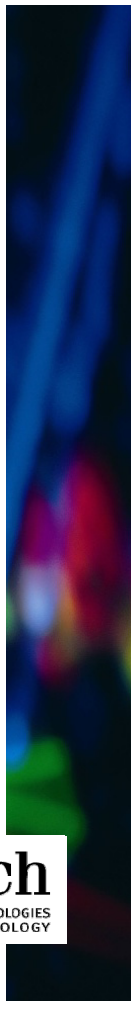
Mots clés : Fouille de données, extraction de connaissances, objets mobiles, surveillance maritime, analyse de comportements.

Methodology of spatio-temporal knowledge discovery through data mining for risk behavior analysis: application to maritime traffic monitoring

ABSTRACT:

The advent of positioning system technologies (AIS, radar, GPS, RFID, etc.), remote transmission (VHF, satellite, GSM, etc.), technological advances in embedded systems and low cost production, has enabled their deployment on a large scale. A huge amount of moving objects data are collected through these technologies and used in various applications such as real time monitoring surveillance of maritime traffic. The post-hoc analysis of data from moving ships and risk events may present interesting opportunities for the understanding and modeling support of risky behaviors. In this work, we propose a methodology based on Spatio-Temporal Data Mining for the knowledge discovery about potentially risky behaviors of ships. Based on this methodology, a workshop to support the analysis of behavior of ships is also proposed.

Keywords : Data mining, Knowledge Discovery, moving objects, maritime monitoring, behavior analysis.



MINES ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY